IEEE Open Journal of Signal Processing

Received 12 February 2025; revised 3 June 2025; accepted 8 June 2025. Date of publication 11 June 2025; date of current version 25 June 2025. The review of this article was arranged by Associate Editor Ajit Rajwade.

Digital Object Identifier 10.1109/OJSP.2025.3578807

On Conditional Independence Graph Learning From Multi-Attribute Gaussian Dependent Time Series

JITENDRA K. TUGNAIT (Life Fellow, IEEE)

Department of Electrical & Computer Engineering, Auburn University, Auburn, AL 36849 USA

(e-mail: tugnajk@auburn.edu)

This work was supported by the National Science Foundation under Grant CCF-2308473.

ABSTRACT Estimation of the conditional independence graph (CIG) of high-dimensional multivariate Gaussian time series from multi-attribute data is considered. Existing methods for graph estimation for such data are based on single-attribute models where one associates a scalar time series with each node. In multi-attribute graphical models, each node represents a random vector or vector time series. In this paper we provide a unified theoretical analysis of multi-attribute graph learning for dependent time series using a penalized log-likelihood objective function formulated in the frequency domain using the discrete Fourier transform of the time-domain data. We consider both convex (sparse-group lasso) and non-convex (log-sum and SCAD group penalties) penalty/regularization functions. We establish sufficient conditions in a high-dimensional setting for consistency (convergence of the inverse power spectral density to true value in the Frobenius norm), local convexity when using non-convex penalties, and graph recovery. We do not impose any incoherence or irrepresentability condition for our convergence results. We also empirically investigate selection of the tuning parameters based on the Bayesian information criterion, and illustrate our approach using numerical examples utilizing both synthetic and real data.

INDEX TERMS Graph estimation, inverse spectral density estimation, multi-attribute data, sparse graph learning, time series, undirected graph.

I. INTRODUCTION

Graphical models are a useful tool for analyzing multivariate data where conditional independence plays an important role [1], [2], [3], [4]. Let $\mathcal{G} = (V, \mathcal{E})$ denote a graph with a set of p vertices (nodes) $V = \{1, 2, \ldots, p\} = [p]$, and a corresponding set of (undirected) edges $\mathcal{E} \subseteq [p] \times [p]$. Consider a stationary, zero-mean, $p-$dimensional multivariate Gaussian time series $\boldsymbol{x}(t)$, $t = 0, \pm 1, \pm 2, \ldots$, with ith component $x_i(t)$, and correlation (covariance) matrix function $\boldsymbol{R}_{xx}(\tau) = \mathbb{E}\{\boldsymbol{x}(t+\tau)\boldsymbol{x}^T(t)\}$, $\tau = 0, \pm 1, \ldots$. Given $\{\boldsymbol{x}(t)\}$, in the corresponding graph $\mathcal{G}$, each component series $\{x_i(t)\}$ is represented by a node (i in V), and associations between components $\{x_i(t)\}$ and $\{x_j(t)\}$ are represented by edges between nodes i and j of $\mathcal{G}$. In a conditional independence graph (CIG), there is no edge between nodes i and j (i.e., $\{i, j\} \notin \mathcal{E}$) if and only if (iff) $x_i(t)$ and $x_j(t)$ are conditionally independent given the remaining p-2 scalar series $x_\ell(t)$, $\ell \in [p]$, $\ell \neq i$, $\ell \neq j$. (This is a generalization of the CIG for random vectors where $\{i, j\} \notin \mathcal{E}$ iff $[\boldsymbol{\Omega}]_{ij} = 0$ ($\boldsymbol{\Omega} = (E\{\boldsymbol{x}(t)\boldsymbol{x}^\top(t)\})^{-1}$) [4], [5], [6].)

Denote the power spectral density (PSD) matrix of $\{\boldsymbol{x}(t)\}$ by $\boldsymbol{S}_x(f)$, where $\boldsymbol{S}_x(f) = \sum_{\tau=-\infty}^{\infty} \boldsymbol{R}_{xx}(\tau) e^{-\imath 2\pi f \tau}$ and $\imath = \sqrt{-1}$. In [6] it was shown that conditional independence of two time series components given all other components of the time series, is encoded by zeros in the inverse PSD, that is, $\{i, j\} \notin \mathcal{E}$ iff the (i, j)-th element of $\boldsymbol{S}_x^{-1}(f)$, $[\boldsymbol{S}_x^{-1}(f)]_{ij} = 0$ for every f. Hence one can use estimated inverse PSD of observed time series to infer the associated graph. In [6] the low-dimensional case is addressed. Nonparametric frequency-domain approaches for graph estimation in high-dimensional settings (sample size n is less than or of the order of p) have been considered in [7] using a group-lasso penalty, and in [8], [9], [10] using a sparse-group lasso penalty. The focus of this paper is on high-dimensional settings where the number of graph nodes p (e.g., time series dimension) is smaller than or

© 2025 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

comparable to the data sample size n [11]. In particular, in a high-dimensional setting, as $n \to \infty$, $\frac{p}{n} \to c > 0$ for some constant c, instead of $\frac{p}{n} \to 0$ as in classical low-dimensional statistical analysis framework [11, Chapter 1]. Such models for the i.i.d. $\{\boldsymbol{x}(t)\}$ case have been extensively studied [4], [5], [11]. If $\frac{p}{n} \ll 1$, we use the term low-dimensional for such cases in this paper. A sparse-group non-convex log-sum penalty is investigated in [12] to regularize the problem considered in [10], motivated by [13]. Refs. [7], [9], [10] provide performance analysis and guarantees.

Parametric modeling (autoregressive (AR) or autoregressive moving average (ARMA) models) based approaches in low-dimensional settings for CIG estimation for time series are discussed in [14], [15], [16], [17], among others. These papers are focused on algorithm development and they do not provide any performance guarantees (such as [10, Theorem 1] or Theorem 1 in this paper). Compared with this paper or [10] where the high-dimensional case is considered, [14], [15], [16], [17] consider a low-dimensional setting. For instance, in the simulation example 1 of [17], one has a 10-dimensional ARMA model implying a 10-node graph ($p = 10$ in our notation) while the data sample size used to illustrate the performance of their algorithm is 1024 ($n = 1024$ in our notation), leading to $p/n = 0.0098 \ll 1$. In contrast, in the synthetic data example in [10, Sec. 6.1], one has $p = 128$ and $n \in \{128, 256, 512, 1024, 2048\}$, leading to $p/n \in \{1, 0.5, 0.25, 0.125, 0.0625\}$. Statistical analysis in the high-dimensional case requires a different set of analytical tools [3], [11]. Estimation of ARMA models for high-dimensional Gaussian time series without considering graphical modeling aspects is discussed in [18] where in [18, Example 1], one has $p = 200$ with varying values of sample size n with some values of $n < 50$.

In many applications, there may be more than one random variable (or scalar time series) associated with a node. This class of graphical models has been called multi-attribute graphical models in [19], [20] where a high-dimensional setting is considered, and vector graphs or networks in [21], [22], [23] where a low-dimensional setting is considered. In a gene regulatory network, one may have different molecular profiles available for a single gene, such as protein, DNA and RNA. Since these molecular profiles are on the same set of biological samples, they constitute multi-attribute data for gene regulatory graphical models in [19]. The motivation for vector graphical models considered in [21], [22], [23] is network analysis for human fMRI data. In this paper in Section VI-B, we model air-quality and meteorological data acquired at different monitoring stations in Beijing [24], [25] as multi-attribute data, with measurements of each variable at m stations modeled as m attributes. Such graphical models have been considered in the literature only for random vectors (i.e., observations originate from an i.i.d. random sequence), not for time series graphical models. The objective of this paper is to fill this gap. Additionally, we consider both convex (sparse-group lasso [26], [27]) and non-convex (log-sum [13] and Smoothly Clipped Absolute Deviation (SCAD) [28], [29]) penalty functions. It is well-known that use of non-convex penalties can yield more accurate results compared to the lasso penalty, i.e., they can produce sparse set of solution like lasso, and approximately unbiased coefficients for large coefficients, unlike lasso [13], [28], [29]. This motivates consideration of the SCAD and log-sum penalties (in addition to the lasso penalty) in this paper. As noted earlier, a sparse-group non-convex log-sum penalty is investigated in [12] to regularize the single-attribute problem considered in [10] where it is shown empirically that the log-sum penalty significantly outperforms the lasso penalty. Hence the interest in non-convex penalties in this paper.

A. RELATED WORK

There appears to be no prior reported work on graphical modeling for multi-attribute dependent time series in high-dimensional settings. Prior work on graphical modeling for single-attribute dependent time series in low-dimensional settings is concerned with testing whether $\{i, j\} \in \mathcal{E}$ for all possible edges in the graph, based on some nonparametric frequency-domain test statistic such as partial coherence [6], [30], [31], [32], [33] which requires estimates of $\boldsymbol{S}_x(f)$. These approaches do not scale to high dimensions where p is comparable to or larger than the sample size n. As an alternative to nonparametric modeling of time series, parametric graphical models utilizing (Gaussian) vector AR (VAR) process models of $\boldsymbol{x}(t)$ have been proposed in [16], [34], [35], [36] and ARMA process (and related) models may be found in [14], [15], [17], but these approaches are suitable only for low-dimensional settings as discussed earlier. These approaches do not address the multi-attribute case. Graphical modeling for single-attribute dependent time series in high-dimensional settings has been considered using nonparametric frequency-domain approaches in [7], [8], [9], [10] with convex lasso-related regularization and in [12] with non-convex log-sum regularization. A time-domain approach with log-sum penalty may be found in [37].

Multi-attribute graphical modeling in high-dimensional setting given i.i.d. data has been addressed in [19], [20] using convex lasso-related regularization and in [38] using non-convex SCAD penalty. When convex regularization is used, the overall optimization problem is convex where a global optimum solution is guaranteed, whereas with non-convex penalties, one can obtain only a local optimum.

This paper builds on the work reported in [10]. A detailed comparison between this paper and [10] (also [12]) is given later in Remark 4 in Section V after we have introduced all the technical details facilitating the comparison.

B. OUR CONTRIBUTIONS

In this paper we provide a unified theoretical analysis of multi-attribute graph learning for dependent time series using a penalized log-likelihood objective function in the frequency domain. We consider the convex sparse-group lasso as well as the non-convex log-sum and SCAD group penalties. The non-convex optimization problem (when using non-convex

penalties) is solved via iterative convex optimization, based on a local-linear approximation (LLA) [29], [39] to the non-convex penalty and an alternating direction method of multipliers (ADMM) method. The ADMM method used in this paper follows [10] and differences between [10] and this paper are explained later in Section IV and in Remark 4 in Section V. We establish sufficient conditions in a high-dimensional setting for consistency (convergence of the inverse power spectral density to true value in the Frobenius norm) in Theorem 1, local convexity when using non-convex penalties in Theorem 2, and graph recovery in Theorem 3. We do not impose any incoherence or irrepresentability condition for our Theorems 1–3 (see Remark 3 in Section V). We illustrate our approach using numerical examples utilizing both synthetic and real (Beijing air-quality [24], [25]) data.

A preliminary version of this paper appears in a conference paper [40] where proofs of Theorems 1 and 3 and Lemma 1 are not given, and only a sketch of proof of Theorem 2 appears. Theorem 1 in [40] has an error. Synthetic data examples are different in this paper and [40].

C. OUTLINE AND NOTATION

The rest of the paper is organized as follows. The underlying system model and the resulting log-likelihood formulation of the problem are presented in Section II. The convex and non-convex penalty functions and their properties (based on [41]), and the resulting penalized negative log-likelihood function is discussed in Section III. A solution to the non-convex optimization problem is provided in Section IV. Selection of the tuning parameters based on BIC is presented in Section IV-A. In Section V we provide a theoretical analysis of the proposed approach, resulting in Theorems 1–3. Numerical results are presented in Section VI and proofs of Theorems 1, 2 and 3 are given in the two appendices.

The superscripts $*$, $\top$ and H denote the complex conjugate, transpose and Hermitian (conjugate transpose) operations, respectively, and the sets of real, positive real and complex numbers are denoted by $\mathbb{R}$, $\mathbb{R}_+$ and $\mathbb{C}$, respectively. Given $\boldsymbol{A} \in \mathbb{C}^{p\times p}$, we use $\phi_{\min}(\boldsymbol{A})$, $\phi_{\max}(\boldsymbol{A})$, $|\mathbf{A}|$, $\mathrm{tr}(\boldsymbol{A})$ and $\mathrm{etr}(\boldsymbol{A})$ to denote the minimum eigenvalue, maximum eigenvalue, determinant, trace, and exponential of trace of $\boldsymbol{A}$, respectively. We use $\boldsymbol{A} \succeq 0$ and $\boldsymbol{A} \succ 0$ to denote that Hermitian $\boldsymbol{A}$ is positive semi-definite and positive definite, respectively, and $\boldsymbol{I}_p$ is the $p \times p$ identity matrix. For $\boldsymbol{B} \in \mathbb{C}^{p\times q}$, we define the operator norm, the Frobenius norm and the vectorized ℓ_1 norm, respectively, as $\|\boldsymbol{B}\| = \sqrt{\phi_{\max}(\boldsymbol{B}^H\boldsymbol{B})}$, $\|\boldsymbol{B}\|_F = \sqrt{\mathrm{tr}(\boldsymbol{B}^H\boldsymbol{B})}$ and $\|\boldsymbol{B}\|_1 = \sum_{i,j} |B_{ij}|$, where B_{ij} is the (i, j)-th element of $\boldsymbol{B}$, also denoted by $[\boldsymbol{B}]_{ij}$. For vector $\boldsymbol{\theta} \in \mathbb{C}^p$, we define $\|\boldsymbol{\theta}\|_1 = \sum_{i=1}^p |\theta_i|$ and $\|\boldsymbol{\theta}\|_2 = \sqrt{\sum_{i=1}^p |\theta_i|^2}$, and we also use $\|\boldsymbol{\theta}\|$ for $\|\boldsymbol{\theta}\|_2$. The Kronecker product of matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ is denotes by $\boldsymbol{A} \otimes \boldsymbol{B}$. Given $\boldsymbol{A} \in \mathbb{C}^{p\times p}$, $\boldsymbol{A}^+ = \mathrm{diag}(\boldsymbol{A})$ is a diagonal matrix with the same diagonal as $\boldsymbol{A}$, and $\boldsymbol{A}^- = \boldsymbol{A} - \boldsymbol{A}^+$ is $\boldsymbol{A}$ with all its diagonal elements set to zero. Given $\boldsymbol{A} \in \mathbb{C}^{n\times p}$, column vector $\mathrm{vec}(\boldsymbol{A}) \in \mathbb{C}^{np}$ denotes the vectorization of $\boldsymbol{A}$ which stacks the columns of the matrix $\boldsymbol{A}$. The notation $\boldsymbol{x} \sim \mathcal{N}_c(\mathbf{m}, \boldsymbol{\Sigma})$ denotes a complex random vector $\boldsymbol{x}$ that is circularly symmetric (proper), complex Gaussian with mean $\boldsymbol{m}$ and covariance $\boldsymbol{\Sigma}$, and $\boldsymbol{x} \sim \mathcal{N}_r(\mathbf{m}, \boldsymbol{\Sigma})$ denotes real-valued Gaussian $\boldsymbol{x}$ with mean $\boldsymbol{m}$ and covariance $\boldsymbol{\Sigma}$.

II. SYSTEM MODEL

Consider p jointly Gaussian, zero-mean stationary, vector sequences $\{z_i(t)\}_{t\in\mathbb{Z}}$, $z_i(t) \in \mathbb{R}^m$, $i \in [p]$. In a multi-attribute time series graphical model, we associate $\{z_i(t)\}_{t\in\mathbb{Z}}$ with the ith node of an undirected graph $\mathcal{G} = (V, \mathcal{E})$ where $V = [p]$ is the set of p nodes (vertices) and $\mathcal{E} \subseteq V \times V$ is the set of undirected edges that describe the conditional dependencies among the p sequences $\{\{z_i(t)\}_{t\in\mathbb{Z}},\ i \in V\}$. Similar to the scalar case ($m = 1$), edge $\{i, j\} \notin \mathcal{E}$ iff the sequences $\{z_i(t)\}$ and $\{z_j(t)\}$ are conditionally independent given the remaining $p - 2$ vector sequences $\{z_\ell(t)\}$, $\ell \in V\backslash\{i, j\}$.

Define the mp-dimensional sequence

$$\boldsymbol{x}(t) = \left[z_1^\top(t),\ z_2^\top(t),\ \cdots,\ z_m^\top(t)\right]^\top \in \mathbb{R}^{mp}. \tag{1}$$

Associate $\{\boldsymbol{x}(t)\}_{t\in\mathbb{Z}}$ with an enlarged graph $\bar{\mathcal{G}} = (\bar{V}, \bar{\mathcal{E}})$ where $\bar{V} = [mp]$ and $\bar{\mathcal{E}} \subseteq \bar{V} \times \bar{V}$. The ℓth component of $\{z_j(t)\}$, denoted by $\{[z_j]_\ell(t)\}$, associated with the node j of $\mathcal{G}$, is the scalar sequence $\{x_q(t)\}$, $x_q = [\boldsymbol{x}]_q$, $q = (j-1)m + \ell$, $j \in [p]$ and $\ell \in [m]$. The scalar sequence $\{x_q(t)\}$ is associated with node q of enlarged graph $\bar{\mathcal{G}}$. Corresponding to the edge $\{j, k\} \in V \times V$ in $\mathcal{G}$, there are m^2 edges $\{q, r\} \in \bar{V} \times \bar{V}$ in $\bar{\mathcal{G}}$ where $q = (j-1)m + u$ and $r = (k-1)m + v$ with $u, v \in [m]$.

As in Section I, denote the power spectral density (PSD) matrix of $\{\boldsymbol{x}(t)\}$ by $\boldsymbol{S}_x(f)$. Here f is the normalized frequency, in Hz. Given a matrix $\boldsymbol{A} \in \mathbb{C}^{(mp)\times(mp)}$, we use $\boldsymbol{A}^{(jk)}$ to denote the $m \times m$ submatrix of $\boldsymbol{A}$ whose (u, v)th element is given by

$$[\boldsymbol{A}^{(jk)}]_{uv} = [\boldsymbol{A}]_{(j-1)m+u,(k-1)m+v}, \quad u, v \in [m]. \tag{2}$$

By [6, Theorem 2.4], in the CIG $\mathcal{G} = (V, \mathcal{E})$ of the multi-attribute time series $\{\boldsymbol{x}(t)\}_{t\in\mathbb{Z}}$ originating via (1), we have

$$\{j, k\} \notin \mathcal{E} \quad \Leftrightarrow \quad \left(\boldsymbol{S}_x^{-1}(f)\right)^{(jk)} \equiv \mathbf{0} \tag{3}$$

provided $\boldsymbol{S}_x(f) \succ \mathbf{0}\ \forall f$. (Note that while most of the discussion and all of the numerical results in [6] pertain to scalar time series per node, the theory is shown to apply to vector series per node also.)

A. PROBLEM FORMULATION

We observe a finite-duration segment $\{\boldsymbol{x}(t)\}_{t=0}^{n-1}$ of a realization of an $mp-$dimensional stationary Gaussian sequence $\{\boldsymbol{x}(t)\}_{t\in\mathbb{Z}}$. Our objective is to first estimate the inverse PSD $\boldsymbol{S}_x^{-1}(f)$ at distinct frequencies, and then select the edge $\{j, k\}$ in the graphical model $\mathcal{G}$ based on whether or not $(\boldsymbol{S}_x^{-1}(f))^{(jk)} = \mathbf{0}$ for every f. The single attribute case ($m = 1$) has been discussed in [10] with sparse-group lasso penalty and in [12] with sparse-group log-sum penalty. Since for a real-valued time series, $\boldsymbol{S}_x(f) = \boldsymbol{S}_x^H(-f)$, and $\boldsymbol{S}_x(f)$ is periodic in f with period one, knowledge of $\boldsymbol{S}_x(f)$ in the interval [0,0.5] completely specifies $\boldsymbol{S}_x(f)$ for other values of

f. Hence, it is enough to check if $(\boldsymbol{S}_x^{-1}(f))^{(jk)} = \mathbf{0}$ for every $f \in [0, 0.5]$.

Given $\{\boldsymbol{x}(t)\}_{t=0}^{n-1}$, define the (normalized) DFT $\boldsymbol{d}_x(f_\ell)$ of $\boldsymbol{x}(t)$, $(\imath = \sqrt{-1})$,

$$\boldsymbol{d}_x(f_\ell) = \frac{1}{\sqrt{n}} \sum_{t=0}^{n-1} \boldsymbol{x}(t) \exp\left(-\imath 2\pi f_\ell t\right), \tag{4}$$

where

$$f_\ell = \ell/n, \quad \ell = 0, 1, \ldots, n-1. \tag{5}$$

Since $\{\boldsymbol{x}(t)\}$ is Gaussian, so is $\boldsymbol{d}_x(f_\ell)$. As discussed in [10], the set of complex-valued random vectors $\{\boldsymbol{d}_x(f_\ell)\}_{\ell=0}^{n/2}$, n even, is a sufficient statistic for any statistical inference problem, including our problem of estimation of inverse PSD.

We need the following assumption in order to invoke [42, Theorem 4.4.1], used extensively later.

A1) The $mp-$dimensional time series $\{\boldsymbol{x}(t)\}_{t\in\mathbb{Z}}$ is zero-mean stationary and Gaussian, satisfying

$$\sum_{\tau=-\infty}^{\infty} |[\boldsymbol{R}_{xx}(\tau)]_{k\ell}| < \infty \text{ for every } k, \ell \in \bar{V}.$$

It follows from [42, Theorem 4.4.1] that under assumption (A1), asymptotically (as $n \to \infty$), $\boldsymbol{d}_x(f_\ell)$, $\ell \in [(n/2) - 1]$, (n even), are independent proper (i.e., circularly symmetric), complex Gaussian $\mathcal{N}_c(\mathbf{0}, \boldsymbol{S}_x(f_\ell))$ random vectors, respectively. Also, asymptotically, $\boldsymbol{d}_x(f_0)$ and $\boldsymbol{d}_x(f_{n/2})$, ($n$ even), are independent real Gaussian $\mathcal{N}_r(\mathbf{0}, \boldsymbol{S}_x(f_0))$ and $\mathcal{N}_r(\mathbf{0}, \boldsymbol{S}_x(f_{n/2}))$ random vectors, respectively, independent of $\boldsymbol{d}_x(f_\ell)$, $\ell \in [(n/2) - 1]$. We will ignore these two frequency points f_0 and $f_{n/2}$.

Define

$$\boldsymbol{D} = \left[\boldsymbol{d}_x(f_1) \ \cdots \ \boldsymbol{d}_x(f_{(n/2)-1})\right] \in \mathbb{C}^{(mp)\times((n/2)-1)}. \tag{6}$$

We assume that $\boldsymbol{S}_x(f_\ell)$ is locally smooth (a standard assumption in PSD estimation [42]), so that $\boldsymbol{S}_x(f_\ell)$ is (approximately) constant over $K = 2m_t + 1$ consecutive frequency points. Pick

$$\tilde{f}_k = \frac{(k-1)K + m_t + 1}{n}, \qquad k = 1, 2, \ldots, M, \tag{7}$$

$$M = \left\lfloor \left(\frac{n}{2} - m_t - 1\right)/K \right\rfloor, \tag{8}$$

leading to M equally spaced frequencies $\tilde{f}_k$ in the interval (0,0.5), at intervals of K/n. We state the local smoothness assumption as assumption (A2).

A2) Assume that for $\ell = -m_t, -m_t + 1, \ldots, m_t$,

$$\boldsymbol{S}_x(\tilde{f}_{k,\ell}) = \boldsymbol{S}_x(\tilde{f}_k), \tag{9}$$

$$\text{where } \tilde{f}_{k,\ell} = ((k-1)K + m_t + 1 + \ell)/n. \tag{10}$$

Under assumptions (A1)-(A2), the joint pdf of $\boldsymbol{D}$ is

$$f_{\boldsymbol{D}}(\boldsymbol{D}) = \prod_{k=1}^{M} \left[\prod_{\ell=-m_t}^{m_t} \frac{\exp\left(-g_{kl} - g_{kl}^*\right)}{\pi^{mp} |\boldsymbol{S}_x^{-1}(\tilde{f}_k)|^{1/2} |\boldsymbol{S}_x^{-*}(\tilde{f}_k)|^{1/2}} \right], \tag{11}$$

$$g_{kl} = \frac{1}{2} \boldsymbol{d}_x^H(\tilde{f}_{k,\ell}) \boldsymbol{S}_x^{-1}(\tilde{f}_k) \boldsymbol{d}_x(\tilde{f}_{k,\ell}), \tag{12}$$

where $\boldsymbol{A}^{-*}$ stands for $(\boldsymbol{A}^{-1})^*$. Parametrizing in terms of the inverse PSD matrix $\boldsymbol{\Phi}_k := \boldsymbol{S}_x^{-1}(\tilde{f}_k)$, the negative log-likelihood, up to some irrelevant constants, is given by

$$-\ln f_{\boldsymbol{D}}(\boldsymbol{D}) \propto \mathcal{L}(\boldsymbol{\Omega}) \tag{13}$$

$$:= \sum_{k=1}^{M} \frac{1}{2} \left[-\ln(|\boldsymbol{\Phi}_k|) - \ln(|\boldsymbol{\Phi}_k^*|) + \operatorname{tr}\left(\hat{\boldsymbol{S}}_k \boldsymbol{\Phi}_k + \hat{\boldsymbol{S}}_k^* \boldsymbol{\Phi}_k^*\right) \right] \tag{14}$$

where

$$\boldsymbol{\Omega} = [\boldsymbol{\Phi}_1, \ \boldsymbol{\Phi}_2, \ \cdots, \ \boldsymbol{\Phi}_M] \in \mathbb{C}^{(mp)\times(mpM)}, \tag{15}$$

$$\hat{\boldsymbol{S}}_k = \frac{1}{K} \sum_{\ell=-m_t}^{m_t} \mathbf{d}_x(\tilde{f}_{k,\ell}) \mathbf{d}_x^H(\tilde{f}_{k,\ell}). \tag{16}$$

Note that $\hat{\boldsymbol{S}}_k$ represents PSD estimator at frequency $\tilde{f}_k$ using unweighted frequency-domain smoothing [42].

Our objective is to estimate $\boldsymbol{\Omega}$ given $\{\boldsymbol{x}(t)\}_{t=0}^{n-1}$, and to infer the underlying CIG based on estimated $\boldsymbol{\Omega}$.

III. PENALIZED NEGATIVE LOG-LIKELIHOOD

To enforce sparsity and to make the problem well-conditioned (when $K < p$), as in [10], we propose to minimize a penalized version $\bar{\mathcal{L}}(\boldsymbol{\Omega})$ of $\mathcal{L}(\boldsymbol{\Omega})$ where we penalize (regularize) at both element-wise and group-wise. We have

$$\bar{\mathcal{L}}(\boldsymbol{\Omega}) = \mathcal{L}(\boldsymbol{\Omega}) + \alpha P_e(\boldsymbol{\Omega}) + (1-\alpha) P_g(\boldsymbol{\Omega}), \tag{17}$$

$$P_e(\boldsymbol{\Omega}) = \sum_{k=1}^{M} \sum_{i\neq j}^{mp} \rho_\lambda\left([\boldsymbol{\Phi}_k]_{ij}\right), \tag{18}$$

$$P_g(\boldsymbol{\Omega}) = m\sqrt{M} \sum_{q\neq\ell}^{p} \rho_\lambda\left(\|\boldsymbol{\Omega}^{(q\ell M)}\|_F\right) \tag{19}$$

where $\boldsymbol{\Omega}^{(q\ell M)} \in \mathbb{C}^{m\times(mM)}$ is defined as

$$\boldsymbol{\Omega}^{(q\ell M)} := \left[\boldsymbol{\Phi}_1^{(q\ell)}, \ \boldsymbol{\Phi}_2^{(q\ell)}, \ \cdots, \ \boldsymbol{\Phi}_M^{(q\ell)}\right], \tag{20}$$

$\boldsymbol{\Phi}_i^{(q\ell)}$, $i \in [M]$, is defined as in (2), $\lambda > 0$, $\alpha \in [0, 1]$, $m\sqrt{M}$ in (19) reflects the number of group variables [43], and for $u \in \mathbb{R}$, $\rho_\lambda(u)$ is a penalty function that is function of $|u|$. In (18), the penalty term is applied to each off-diagonal element of $\boldsymbol{\Phi}_k$ and in (19), the penalty term is applied to the off-block-diagonal group of $m^2 M$ terms via $\boldsymbol{\Omega}^{(q\ell M)}$, defined in (20). The parameter $\alpha \in [0, 1]$ "balances" element-wise and group-wise penalties [10], [26]

The following penalty functions are considered:

- *Lasso*. For some $\lambda > 0$,

$$\rho_\lambda(u) = \lambda|u|, \quad u \in \mathbb{R}\,. \tag{21}$$

- *Log-sum*. For some $\lambda > 0$ and $1 \gg \epsilon > 0$,

$$\rho_\lambda(u) = \lambda\epsilon \ln\left(1 + \frac{|u|}{\epsilon}\right). \tag{22}$$

- *Smoothly Clipped Absolute Deviation (SCAD)*. For some $\lambda > 0$ and $a > 2$,

$$\rho_\lambda(u) = \begin{cases} \lambda|u| & \text{for } |u| \le \lambda \\ \frac{2a\lambda|u| - |u|^2 - \lambda^2}{2(a-1)} & \text{for } \lambda < |u| < a\lambda \\ \frac{\lambda^2(a+1)}{2} & \text{for } |u| \ge a\lambda\,. \end{cases} \tag{23}$$

In the terminology of [41], all of the above three penalties are "μ-amenable" for some $\mu \ge 0$. As defined in [41, Sec. 2.2], $\rho_\lambda(u)$ is μ-amenable for some $\mu \ge 0$ if

i) The function $\rho_\lambda(u)$ is symmetric around zero, i.e., $\rho_\lambda(u) = \rho_\lambda(-u)$ and $\rho_\lambda(0) = 0$.
ii) The function $\rho_\lambda(u)$ is non-decreasing on $\mathbb{R}_+$.
iii) The function $\rho_\lambda(u)/u$ is non-increasing on $\mathbb{R}_+$.
iv) The function $\rho_\lambda(u)$ is differentiable for $u \ne 0$.
v) The function $\rho_\lambda(u) + \frac{\mu}{2}u^2$ is convex, for some $\mu \ge 0$.
vi) $\lim_{u\to 0^+} \rho'(u) = \lambda$ where $\rho'(u) := \frac{d\rho_\lambda(u)}{du}$.

It is shown in [41, Appendix A], that all of the above three penalties are μ-amenable with $\mu = 0$ for Lasso and $\mu = 1/(a-1)$ for SCAD. In [41] the log-sum penalty is defined as $\rho_\lambda(u) = \ln(1 + \lambda|u|)$ whereas in [13], it is defined as $\rho_\lambda(u) = \lambda \ln(1 + \frac{|u|}{\epsilon})$. We follow [13] but modify it so that property (vi) in the definition of μ-amenable penalties holds. In our case $\mu = \frac{\lambda}{\epsilon}$ for the log-sum penalty since $\frac{d^2\rho_\lambda(u)}{du^2} = -\lambda\epsilon/(\epsilon + |u|)^2$ for $u \ne 0$.

The above three penalty functions also have the following properties:

vii) For some $C_\lambda > 0$ and $\delta_\lambda > 0$, the function $\rho_\lambda(u)$ has a lower bound

$$\rho_\lambda(u) \ge C_\lambda|u| \text{ for } |u| \le \delta_\lambda\,. \tag{24}$$

viii) $\frac{d\rho_\lambda(u)}{d|u|} \le \lambda$ for $u \ne 0$.

Property (viii) is straightforward to verify. For Lasso, $C_\lambda = \lambda$ and $\delta_\lambda = \infty$. For SCAD, $C_\lambda = \lambda$ and $\delta_\lambda = \lambda$. Since $\ln(1 + x) \ge x/(1 + x)$ for $x > -1$, we have $\ln(1 + x) \ge x/C_1$ for $0 \le x \le C_1 - 1, C_1 > 1$. Take $C_1 = 2$. Then log-sum $\rho_\lambda(u) \ge \frac{\lambda}{2}|u|$ for any $|u| \le \epsilon$, leading to $C_\lambda = \frac{\lambda}{2}$ and $\delta_\lambda = \epsilon$. We may and will take $C_\lambda = \frac{\lambda}{2}$ for lasso and SCAD penalties as well.

IV. OPTIMIZATION

For non-convex $\rho_\lambda(u)$, we will use a local linear approximation (LLA) as in [29], [39], to yield

$$\rho_\lambda(u) \approx \rho_\lambda(|u_0|) + \rho'_\lambda(|u_0|)(|u| - |u_0|) \;\Rightarrow\; \rho'_\lambda(|u_0|)|u|\,, \tag{25}$$

where u_0 is an initial guess, $\rho'_\lambda(|u_0|) = \lambda\epsilon/(|u_0| + \epsilon)$ for LSP, and for SCAD, $\rho'_\lambda(|u_0|) = \lambda$ for $|u| \le \lambda$, $= \frac{a\lambda - |u|}{a-1}$ for $\lambda < |u| < a\lambda$, and $= 0$ for $|u| \ge a\lambda$. Therefore, with u_0 fixed, we consider only the last term above for optimization w.r.t. u. By [39, Theorem 1], the LLA provides a majorization of the non-convex penalty, thereby yielding a majorization-minimization approach. In fact, by [39, Theorem 2], the LLA is the best convex majorization of the LSP and SCAD penalties. Thus in LSP, with some initial guess $\bar{\boldsymbol{\Phi}}_k$, we replace $\rho_\lambda(|[\boldsymbol{\Phi}_k]_{ij}|) \to \lambda\epsilon/(|[\bar{\boldsymbol{\Phi}}_k]_{ij}| + \epsilon) =: \lambda_{kij}$ and $\rho_\lambda(\|\boldsymbol{\Phi}^{(q\ell M)}\|_F) \to \lambda\epsilon/(\|\bar{\boldsymbol{\Phi}}^{(q\ell M)}\|_F + \epsilon) =: \lambda_{q\ell M}$, leading to an adaptive sparse-group lasso convex problem. The initial guess follows from the solution to lasso-penalized objective function. For SCAD, we have $\lambda_{kij} = \lambda$ for $|[\boldsymbol{\Phi}_k]_{ij}| \le \lambda$, $= (a\lambda - |[\boldsymbol{\Phi}_k]_{ij}|)/(a-1)$ for $\lambda < |[\boldsymbol{\Phi}_k]_{ij}| \le a\lambda$, and $= 0$ otherwise, and similarly for $\lambda_{q\ell M}$.

With LLA, the objective function is transformed to

$$\tilde{\mathcal{L}}(\boldsymbol{\Omega}) = \mathcal{L}(\boldsymbol{\Omega}) + \alpha\tilde{P}_e(\boldsymbol{\Omega}) + (1-\alpha)\tilde{P}_g(\boldsymbol{\Omega}), \tag{26}$$

$$\tilde{P}_e(\boldsymbol{\Omega}) = \sum_{k=1}^{M}\sum_{i\ne j}^{mp} \lambda_{kij}\left|[\boldsymbol{\Phi}_k]_{ij}\right|, \tag{27}$$

$$\tilde{P}_g(\boldsymbol{\Omega}) = m\sqrt{M}\sum_{q\ne\ell}^{p} \lambda_{q\ell M}\|\boldsymbol{\Phi}^{(q\ell M)}\|_F\,. \tag{28}$$

For lasso, we have $\lambda_{kij} = \lambda\ \forall k, i, j$ and $\lambda_{q\ell M} = \lambda\ \forall q, \ell$. We follow an ADMM approach, as outlined in [10], for both lasso and LLA to LSP/SCAD. Consider the scaled augmented Lagrangian [50] for this problem after variable splitting, given by

$$\bar{\mathcal{L}}_\rho(\boldsymbol{\Omega}, \{\boldsymbol{W}\}, \{\boldsymbol{U}\}) = \mathcal{L}(\boldsymbol{\Omega}) + \alpha\tilde{P}_e(\boldsymbol{W}) + (1-\alpha)\tilde{P}_g(\boldsymbol{W}) + \frac{\rho}{2}\sum_{k=1}^{M}\|\boldsymbol{\Phi}_k - \boldsymbol{W}_k + \boldsymbol{U}_k\|_F^2\,, \tag{29}$$

where $\{\boldsymbol{W}\} = \{\boldsymbol{W}_k,\ k \in [M]\}$ results from variable splitting where in the penalties we use $\boldsymbol{W}_k$'s instead of $\boldsymbol{\Phi}_k$'s, adding the equality constraint $\boldsymbol{W}_k = \boldsymbol{\Phi}_k$, $\{\boldsymbol{U}\} = \{\boldsymbol{U}_k,\ k \in [M]\}$ are dual variables, and $\rho > 0$ is the "penalty parameter" [50].

The main difference between [10] and this paper is that in [10], $\boldsymbol{W}_k$ and $\boldsymbol{\Phi}_k$ are $p \times p$ whereas in this paper, we have $\boldsymbol{W}_k$ and $\boldsymbol{\Phi}_k$ as $(mp) \times (mp)$ matrices. Therefore, the approach of [10] is applicable after we account for the dimension difference, and additionally, for the fact that $P_g(\boldsymbol{W})$ and $P_g(\boldsymbol{\Omega})$ are penalized slightly differently in the two papers (the factor $m\sqrt{M}$ is missing from [10]). See [10] for further details. For non-convex penalties (not considered in [10]), we have an iterative solution: first solve with lasso penalty, then use the LLA formulation and solve the resulting adaptive lasso type convex problem. In practice, just two iterations seem to be enough. A pseudocode for the ADMM algorithm used in this paper is given in Algorithm 1 where we use the stopping (convergence) criterion following [50, Sec. 3.3.1] and varying penalty parameter ρ following [50, Sec. 3.4.1]. The variables defined in (30)–(36) are needed in Algorithm 1 with $\boldsymbol{\Phi}_k^{(t+1)}$, $\boldsymbol{W}_k^{(t+1)}, \boldsymbol{U}_k^{(t+1)}$ as defined therein:

$$e_1 = \|[\boldsymbol{\Phi}_1^{(t+1)}, \ldots, \boldsymbol{\Phi}_M^{(t+1)}]\|_F \tag{30}$$

Algorithm 1: ADMM Algorithm for Solving (26)–(29)

Input: PSD estimator $\hat{\boldsymbol{S}}_k$, $k \in [M]$ (computed using (4) and (16)), regularization and penalty parameters λ_{kij} ($i, j \in [mp]$, $k \in [M]$), $\lambda_{q\ell M}$ ($q, \ell \in [p]$), α and $\rho = \bar{\rho}$, tolerances τ_{abs} and τ_{rel}, variable penalty factor $\bar{\mu}$, maximum number of iterations t_{max}. Initial guess $\bar{\boldsymbol{\Phi}}_k$, $k \in [M]$.

Output: Estimated $\hat{\boldsymbol{\Phi}}_k$, $k \in [M]$, and edge-set $\hat{\mathcal{E}}$

1: Initialize: $\boldsymbol{U}_k^{(0)} = \boldsymbol{W}_k^{(0)} = \boldsymbol{0}$, $\boldsymbol{\Phi}_k^{(0)} = \bar{\boldsymbol{\Phi}}_k$, $\rho^{(0)} = \bar{\rho}$

2: converged = **false**, $t = 0$

3: **while** converged = **false** **and** $t \leq t_{max}$, **do**

4: Let $\boldsymbol{V}_k \boldsymbol{J}_k \boldsymbol{V}_k^H$ denote the eigen-decomposition of Hermitian $\hat{\boldsymbol{S}}_k - \rho^{(t)}(\boldsymbol{W}_k^{(t)} - \boldsymbol{U}_k^{(t)})$, $k \in [M]$, with the diagonal matrix $\boldsymbol{J}_k$ consisting of its eigenvalues. Define a diagonal matrix $\tilde{\boldsymbol{J}}_k$ with ℓth diagonal element $\tilde{\boldsymbol{J}}_{k\ell\ell} = \left(-\boldsymbol{J}_{k\ell\ell} + \sqrt{\boldsymbol{J}_{k\ell\ell}^2 + 4\rho^{(t)}}\right)/(2\rho^{(t)})$ where $\boldsymbol{J}_{k\ell\ell} = [\boldsymbol{J}_k]_{\ell\ell}$. Set $\boldsymbol{\Phi}_k^{(t+1)} = \boldsymbol{V}_k \tilde{\boldsymbol{J}}_k \boldsymbol{V}_k^H$.

5: Define soft thresholding scalar operator $T_{st}(a, \beta) := (1 - \beta/|a|)_+ a$ and elementwise matrix soft thresholding operator $\boldsymbol{T}_{st}(\boldsymbol{A}, \alpha)$, specified by $[\boldsymbol{T}_{st}(\boldsymbol{A}, \alpha)]_{uv} := T_{st}([\boldsymbol{A}]_{uv}, \alpha)$, where $(a)_+ := \max(0, a)$ and $u, v \in [m]$. For $k \in [M]$, define $\boldsymbol{A}_k = \boldsymbol{\Phi}_k^{(t+1)} + \boldsymbol{U}_k^{(t)}$ and let $(\boldsymbol{A}_k)^{(q\ell)} \in \mathbb{C}^{m \times m}$ be defined as in (2). Then the diagonal subblocks $(\boldsymbol{W}_k)^{(qq)} \in \mathbb{C}^{m \times m}$ of $\boldsymbol{W}_k$ are updated as ($k \in [M]$)

$$[(\boldsymbol{W}_k^{(t+1)})^{(qq)}]_{uv} = \begin{cases} [\boldsymbol{A}_k^{(qq)}]_{uu} & \text{if } u = v \\ T_{st}([\boldsymbol{A}_k^{(qq)}]_{uv}, \frac{\alpha \lambda_{kij}}{\rho^{(t)}}) & \text{if } u \neq v \end{cases}$$

$q \in [p]$, $u, v \in [m]$, $i = (q-1)m + u$, $j = (q-1)m + v$. The off-diagonal $m \times m$ subblocks of $\boldsymbol{W}_k$ are updated as

$$(\boldsymbol{W}_k^{(t+1)})^{(q\ell)} = \boldsymbol{B}\left(1 - \frac{(1-\alpha)m\sqrt{M}\,\lambda_{q\ell M}}{\rho^{(t)}\|\boldsymbol{B}\|_F}\right)_+$$

where $m \times m$ $\boldsymbol{B}$ has its (u, v)th element as $[\boldsymbol{B}]_{uv} = T_{st}([\boldsymbol{A}_k^{(q\ell)}]_{uv}, \alpha\lambda_{kij}/\rho^{(t)})$, $i = (q-1)m + u$, $j = (\ell - 1)m + v$.

6: Dual update $\boldsymbol{U}_k^{(t+1)} = \boldsymbol{U}_k^{(t)} + (\boldsymbol{\Phi}_k^{(t+1)} - \boldsymbol{W}_k^{(t+1)})$, $k \in [M]$.

7: Check convergence. With $e_1, e_2, e_3, \boldsymbol{R}_p^{(t+1)}, \boldsymbol{R}_d^{(t+1)}$, τ_{pri} and τ_{dual} as defined in (30)–(36), respectively, let $d_p = \|\boldsymbol{R}_p^{(t+1)}\|_F$ and $d_d = \|\boldsymbol{R}_d^{(t+1)}\|_F$. If ($d_p \leq \tau_{pri}$) **and** ($d_d \leq \tau_{dual}$), set converged = **true**.

8: Update penalty parameter ρ :

$$\rho^{(t+1)} = \begin{cases} 2\rho^{(t)} & \text{if } d_p > \bar{\mu} d_d \\ \rho^{(t)}/2 & \text{if } d_d > \bar{\mu} d_p \\ \rho^{(t)} & \text{otherwise}. \end{cases}$$

We also need to set $\boldsymbol{U}^{(t+1)} = \boldsymbol{U}^{(t+1)}/2$ for $d_p > \bar{\mu} d_d$ and $\boldsymbol{U}^{(t+1)} = 2\boldsymbol{U}^{(t+1)}$ for $d_d > \bar{\mu} d_p$.

9: $t \leftarrow t + 1$

10: **end while**

11: Denote the converged inverse PSD estimates as $\hat{\boldsymbol{\Phi}}_k$ and let $\hat{\boldsymbol{\Omega}} = [\hat{\boldsymbol{\Phi}}_1, \cdots, \hat{\boldsymbol{\Phi}}_M]$. With $\hat{\boldsymbol{\Omega}}^{(q\ell M)}$ as in (20), for $q \neq \ell$, if $\|\hat{\boldsymbol{\Omega}}^{(q\ell M)}\|_F > 0$, assign edge $\{q, \ell\} \in \hat{\mathcal{E}}$, else $\{q, \ell\} \notin \hat{\mathcal{E}}$.

$$e_2 = \|[\boldsymbol{W}_1^{(t+1)}, \ldots, \boldsymbol{W}_M^{(t+1)}]\|_F \tag{31}$$

$$e_3 = \|[\boldsymbol{U}_1^{(t+1)}, \ldots, \boldsymbol{U}_M^{(t+1)}]\|_F \tag{32}$$

$$\boldsymbol{R}_p^{(t+1)} = \left[\boldsymbol{\Phi}_1^{(t+1)} - \boldsymbol{W}_1^{(t+1)}, \quad \cdots, \quad \boldsymbol{\Phi}_M^{(t+1)} - \boldsymbol{W}_M^{(t+1)}\right] \tag{33}$$

$$\boldsymbol{R}_d^{(t+1)} = \rho^{(t)}\left[\boldsymbol{W}_1^{(t+1)} - \boldsymbol{W}_1^{(t)}, \quad \cdots, \quad \boldsymbol{W}_M^{(t+1)} - \boldsymbol{W}_M^{(t)}\right] \tag{34}$$

$$\tau_{pri} = mp\sqrt{M}\,\tau_{abs} + \tau_{rel}\,\max(e_1, e_2) \tag{35}$$

$$\tau_{dual} = mp\sqrt{M}\,\tau_{abs} + \tau_{rel}\,e_3/\rho^{(t)}. \tag{36}$$

Our ADMM-based optimization algorithm is as follows.

1) Given M and $K = 2m_t + 1$, calculate $\hat{\boldsymbol{S}}_k$. Initialize iteration $\tilde{m} = 1$, $\boldsymbol{\Omega}^{(0)} = \boldsymbol{0}$, $\bar{\boldsymbol{\Omega}} = [\bar{\boldsymbol{\Phi}}_1, \ldots, \bar{\boldsymbol{\Phi}}_M] = \boldsymbol{\Omega}^{(0)}$ and use $\bar{\boldsymbol{\Omega}}$ to compute λ_{kij}'s and $\lambda_{q\ell M}$.
2) Execute Algorithm 1 with initial guess $\bar{\boldsymbol{\Phi}}_k$, $k \in [M]$.
3) Quit if using lasso, else set $\boldsymbol{\Omega}^{(\tilde{m})} = \hat{\boldsymbol{\Omega}}$ and $\bar{\boldsymbol{\Omega}} = \boldsymbol{\Omega}^{(\tilde{m})}$ to re-compute λ_{kij}'s and $\lambda_{q\ell M}$'s via the LLA. Let $\tilde{m} \leftarrow \tilde{m} + 1$.
4) Repeat steps 2 and 3 until convergence. The converged $\hat{\boldsymbol{\Omega}}$ is the final estimate of the inverse PSD's. (For the numerical results shown in Section VI, we terminated after two iterations of steps 2 and 3, similar to [29], [39].)

For the numerical results in Section VI, we used $\bar{\mu} = 10$, $\bar{\rho} = 2$, $\epsilon = 0.0001$ for log-sum penalty, $a = 3.7$ (as in [28], [29]) for the SCAD penalty, $\tau_{abs} = \tau_{rel} = 10^{-4}$ and $t_{\max} = 200$.

A. B.I.C. FOR TUNING PARAMETER SELECTION

Given n and choice of K and M, we follow the Bayesian information criterion (BIC) as given in [10], to select λ (with $\alpha = 0.05$ fixed), for all penalty functions. The Bayesian information criterion (BIC) of [10] is given by

$$\begin{aligned} \text{BIC}(\lambda, \alpha) = 2K \sum_{k=1}^{M} \left(-\ln|\hat{\boldsymbol{\Phi}}_k| + \text{tr}\left(\hat{\boldsymbol{S}}_k \hat{\boldsymbol{\Phi}}_k\right)\right) \\ + \ln(2KM) \sum_{k=1}^{M} (\#\text{ of nonzero elements in } \hat{\boldsymbol{\Phi}}_k). \end{aligned} \tag{37}$$

We use BIC to select λ from over a grid of values. We search over λ in the range $[\lambda_\ell, \lambda_u]$ selected via the following heuristic (similar to [10], [20]). We find the smallest λ, labeled λ_{sm}, for which we get a no-edge model (i.e., $|\hat{\mathcal{E}}| = 0$). Then we

set $\lambda_u = \lambda_{sm}/2$ and $\lambda_\ell = \lambda_u/10$. The given choice of λ_u precludes "extremely" sparse models while that of λ_ℓ precludes "very" dense models.

We note that there exist other general approaches for tuning parameter selection such as cross-validation. Cross-validation generally involves first partitioning the data into K-subsets (K=5 or 10 folds), i.e., K non-overlapping subsets picked randomly. Then $K-1$ segments act as training data for model fitting and the remaining segment is used as test (or validation) set. Tuning parameter would be picked to minimize a test set measure (e.g., negative log-likelihood or some other non-penalized original objective function) after averaging over several partitions. For instance, [27] uses such a method where the data is assumed to be i.i.d. For dependent data with frequency-domain approaches, there are several unresolved issues. For instance, the DFT over training and test datasets would have different resolution since sample size would be different. Moreover, to preserve time dependency, one cannot sample as for i.i.d. data; one must sample block-wise to keep contiguous data-points together. These are unresolved issues which precludes the use of cross-validation in our case.

V. THEORETICAL ANALYSIS

We now allow p, M, K (see (7), (8)), and λ to be functions of sample size n, denoted as p_n, M_n, K_n and λ_n, respectively. We take p_n to be a non-decreasing function of n, as is typical in high-dimensional settings. Note that $K_n M_n \approx n/2$. Pick $K_n = a_1 n^\gamma$ and $M_n = a_2 n^{1-\gamma}$ for some $0.5 < \gamma < 1$, $0 < a_1, a_2 < \infty$, so that both $M_n/K_n \to 0$ and $K_n/n \to 0$ as $n \to \infty$ (cf. [10, Remark 1]).

Recall that we have the original multi-attribute graph $\mathcal{G} = (V, \mathcal{E})$ with $|V| = p_n$ and the enlarged graph $\bar{\mathcal{G}} = (\bar{V}, \bar{\mathcal{E}})$ with $|\bar{V}| = mp_n$. We assume (A3) below regarding $\mathcal{G}$.

(A3) Denote the true edge set of the graph by $\mathcal{E}_0$, implying that $\mathcal{E}_0 = \{\{j,k\} : (\boldsymbol{S}_0^{-1}(f))^{(jk)} \not\equiv 0,\ j \neq k,\ 0 \leq f \leq 0.5\}$ where $\boldsymbol{S}_0(f)$ denotes the true PSD of $\boldsymbol{x}(t)$. (We also use $\boldsymbol{\Phi}_{0k}$ for $\boldsymbol{S}_0^{-1}(\tilde{f}_k)$ where $\tilde{f}_k$ is as in (7), and use $\boldsymbol{\Omega}_0$ to denote the true value of $\boldsymbol{\Omega}$). Assume that $\text{card}(\mathcal{E}_0) = |\mathcal{E}_0| \leq s_{n0}$.

(A4) The minimum and maximum eigenvalues of $mp_n \times mp_n$ PSD $\boldsymbol{S}_0(f) \succ \boldsymbol{0}$ satisfy

$$0 < \beta_{\min} \leq \min_{f\in[0,0.5]} \phi_{\min}(\boldsymbol{S}_0(f)) \leq \max_{f\in[0,0.5]} \phi_{\max}(\boldsymbol{S}_0(f)) \leq \beta_{\max} < \infty\,.$$

Here $\beta_{\min}$ and $\beta_{\max}$ are not functions of n (or p_n).

Let $\hat{\boldsymbol{\Omega}}_\lambda = \arg\min_{\boldsymbol{\Omega}:\boldsymbol{\Phi}_k \succ \boldsymbol{0}} \bar{\mathcal{L}}(\boldsymbol{\Omega})$. Theorem 1 establishes local consistency of $\hat{\boldsymbol{\Omega}}_\lambda$ for non-convex penalties and global consistency for the convex penalty.

Theorem 1: (Local Consistency). For $\tau > 2$, let

$$C_0 = 80 \max_{\ell,f}([\boldsymbol{S}_0(f)]_{\ell\ell})\sqrt{N_0/\ln(mp_n)} \tag{38}$$

where

$$N_0 = 2\ln(16(mp_n)^\tau M_n)\,. \tag{39}$$

Define

$$R = 8(1+m)C_0/\beta_{\min}^2\,, \tag{40}$$

$$r_n = \sqrt{M_n(mp_n + m^2 s_{n0})\ln(mp_n)/K_n} = o(1)\,, \tag{41}$$

$$N_1 = \arg\min\{n : K_n > N_0\}\,, \tag{42}$$

$$N_2 = \arg\min\{n : r_n \leq 0.1/(R\beta_{\min})\}\,, \tag{43}$$

$$N_3 = \arg\min\{n : r_n \leq \epsilon/R\}\,, \tag{44}$$

$$N_4 = \arg\min\left\{n : \lambda_n \leq \frac{\min_{(i,j):[\boldsymbol{\Omega}_0]_{ij}\neq 0}|[\boldsymbol{\Omega}_0]_{ij}|}{a+1}\right\}, \tag{45}$$

$$\lambda_{n\ell} = 2C_0\sqrt{\ln(mp_n)/K_n}\,, \tag{46}$$

$$\lambda_{nu1} = C_0\,\frac{1+m}{m}\sqrt{\left(m^2 + \frac{mp_n}{s_{n0}}\right)\frac{\ln(mp_n)}{K_n}}\,, \tag{47}$$

$$\lambda_{nu2} = \max(Rr_n, \lambda_{nu1})\,. \tag{48}$$

Under assumptions (A1)-(A4), there exists a local minimizer $\hat{\boldsymbol{\Omega}}_\lambda$ of $\bar{\mathcal{L}}(\boldsymbol{\Omega})$ satisfying

$$\|\hat{\boldsymbol{\Omega}}_\lambda - \boldsymbol{\Omega}_0\|_F \leq Rr_n \tag{49}$$

with probability $> 1 - 1/(mp_n)^{\tau-2}$ if

i) for the lasso penalty $\rho_\lambda(t) = \lambda|t|$, sample size $n > \max\{N_1, N_2\}$ and λ_n satisfies $\lambda_{n\ell} \leq \lambda_n \leq \lambda_{nu1}$,

ii) for the SCAD penalty $\rho_\lambda(t)$, sample size $n > \max\{N_1, N_2, N_4\}$ and $\lambda_n = \lambda_{nu2}$,

iii) sample size $n > \max\{N_1, N_2, N_3\}$ and λ_n satisfies $\lambda_{n\ell} \leq \lambda_n \leq \lambda_{nu1}$ for the log-sum penalty $\rho_\lambda(t)$.

For the lasso penalty, $\hat{\boldsymbol{\Omega}}_\lambda$ is a global minimizer whereas for the other two penalties, it is a local minimizer. •

The proof of Theorem 1 is given in Appendix A.

Remark 1: Theorem 1 helps determine how to choose M_n and K_n so that for given p_n, $\lim_{n\to\infty}\|\hat{\boldsymbol{\Omega}}_\lambda - \boldsymbol{\Omega}_0\|_F = 0$ (see also [10, Remark 2]). This behavior is governed by (49), therefore we have to examine r_n. As noted before, since $K_n M_n \approx n/2$, if one picks $K_n = a_1 n^\gamma$, then $M_n = a_2 n^{1-\gamma}$ for some $0 < \gamma < 1$, $0 < a_1, a_2 < \infty$. Suppose that $p_n + ms_{n0}$ satisfies $p_n + ms_{n0} = a_3 n^\theta$ for some $0 \leq \theta < 1, 0 < a_3 < \infty$. Then for fixed m, we have

$$\mathcal{O}(r_n) = \mathcal{O}\left(\frac{(\ln(n))^{1/2} n^{(1-\gamma)/2} n^{\theta/2}}{n^{\gamma/2}}\right) = \mathcal{O}\left(\frac{(\ln(n))^{1/2}}{n^{(2\gamma-1-\theta)/2}}\right) \overset{n\uparrow\infty}{\to} 0 \text{ if } 2\gamma - 1 - \theta > 0\,. \tag{50}$$

Therefore, we must have $1 > \gamma > \frac{1}{2} + \frac{\theta}{2}$. If $\theta = 0$ (fixed graph size and fixed number of connected edges w.r.t. sample size n), we need $\frac{1}{2} < \gamma < 1$. If $\theta > 0$, γ has to be increased beyond what is needed for $\theta = 0$, implying more smoothing of periodogram $\boldsymbol{d}_x(f_m)\boldsymbol{d}_x^H(f_m)$ around f_k to estimate $\boldsymbol{S}_x(f_k)$ (recall (16)), leading to fewer frequency test points M_n. Clearly, we cannot have $\theta \geq 1$ because $p_n + ms_{n0} = \mathcal{O}(n^\theta)$ will require $\gamma > 1$ which is impossible. □

We follow the proof technique of [41, Lemma 6] in establishing Lemma 1 whose proof is in Appendix B. □

Lemma 1. (Local Convexity): The optimization problem

$$\hat{\mathbf{\Omega}}_\lambda = \arg \min_{\mathbf{\Omega}\,:\,\mathbf{\Phi}_k \in \mathcal{B}_k} \bar{\mathcal{L}}(\mathbf{\Omega}), \tag{51}$$

$$\mathcal{B}_k = \left\{ \mathbf{\Phi}_k \,:\, \mathbf{\Phi}_k \succ \mathbf{0},\ \|\mathbf{\Phi}_k\| \le 0.99\sqrt{2/(m\mu\sqrt{M_n})} \right\}, \tag{52}$$

$$\sqrt{2/(m\mu\sqrt{M_n})} = \begin{cases} \infty & : \text{Lasso} \\ \sqrt{\frac{2(a-1)}{m\sqrt{M_n}}} & : \text{SCAD} \\ \sqrt{\frac{2\epsilon}{m\sqrt{M_n}\lambda_n}} & : \text{log-sum}, \end{cases} \tag{53}$$

consists of a strictly convex objective function over a convex constraint set, for all three penalties, where λ_n is as defined in Theorem 1. •

Lemma 1 and Theorem 1 lead to Theorem 2 which is proved in Appendix B.

Theorem 2: Assume the conditions of Theorem 1. Then $\hat{\mathbf{\Omega}}_\lambda$ as defined in Lemma 1 is unique, satisfying $\|\hat{\mathbf{\Omega}}_\lambda - \mathbf{\Omega}_0\|_F \le Rr_n$ with probability $> 1 - 1/(mp_n)^{\tau-2}$ if $Rr_n + 1/\beta_{\min} \le 0.99\sqrt{2/(m\mu\sqrt{M_n})}$, as defined in Lemma 1. •

Remark 2: With lasso, (51) is obviously a globally convex optimization problem since $\mathcal{B}_k = \{\mathbf{\Phi}_k \,:\, \mathbf{\Phi}_k \succ \mathbf{0}\}$, hence, Theorems 1 and 2 yield a unique global optimum. For the SCAD penalty, $\sqrt{\frac{2(a-1)}{m\sqrt{M_n}}} = \mathcal{O}(a^{1/2}/n^{(1-\gamma)/4})$ with M_n as in Remark 1. For fixed SCAD parameter a, with increasing n the convexity region shrinks. To counter this, one could allow a to increase, but this would make SCAD more like lasso. To consider log-sum penalty, using (41) and (47), we express λ_{nu1} as $\lambda_{nu1} = C_0(1+m)r_n/\sqrt{m^2 s_{n0} M_n}$, which together with $\lambda_n \le \lambda_{nu1}$ implies that

$$\sqrt{\frac{2\epsilon}{m\sqrt{M_n}\lambda_n}} \ge \sqrt{\frac{2\epsilon\sqrt{s_{n0}}}{C_0(1+m)r_n}} \overset{n\uparrow\infty}{\rightarrow} \infty. \tag{54}$$

Now with increasing n, the convexity region expands, unlike SCAD. □

We now turn to graph recovery. We follow the proof technique of [45, Theorem 10] in establishing Theorem 3 whose proof is in Appendix B. For some $\gamma_n > 0$, define

$$\hat{\mathcal{E}} = \left\{ \{q,\ell\} \,:\, \left\| \hat{\mathbf{\Omega}}^{(q\ell M_n)} \right\|_F > \gamma_n > 0, q \ne \ell \right\}, \tag{55}$$

$$\mathcal{E}_0 = \left\{ \{q,\ell\} \,:\, \left\| \mathbf{\Omega}_0^{(q\ell M_n)} \right\|_F > 0, q \ne \ell \right\}, \tag{56}$$

$$\bar{\sigma}_n = Rr_n, \tag{57}$$

$$\nu = \min_{\{q,\ell\}\in\mathcal{E}_0} \|\mathbf{\Omega}_0^{(q\ell M_n)}\|_F, \tag{58}$$

$$N_5 = \arg\min\{n \,:\, \bar{\sigma}_n \le 0.4\nu\}, \tag{59}$$

where R and r_n are as in (40) and (41), respectively.

Theorem 3: For $\gamma_n = 0.5\nu$ and $n \ge N_5$, $\hat{\mathcal{E}} = \mathcal{E}_0$ with probability $> 1 - 1/(mp_n)^{\tau-2}$ under the conditions of Theorem 1. □

Remark 3: In practice we do not know the value of ν, hence cannot calculate γ_n needed in (55). For the numerical results presented in Section VI, we used $\gamma_n = 0$. Using some incoherence or irrepresentability conditions and the primal-dual witness method (as in [19], [46]), it may be possible to establish a result similar to Theorem 3 but with $\gamma_n = 0$. We leave this for future research. We do not impose any incoherence or irrepresentability conditions in this paper. □

Remark 4: We now provide a detailed comparison between this paper and [10] (also [12]). The differences between this paper and [10], [12] are as follows.

i) As discussed in Secs. II and IV, in this paper we have $(mp) \times (mp)$ inverse PSD matrices $\mathbf{\Phi}_k$, $k \in [M]$, for a p-node graph, compared to $p \times p$ $\mathbf{\Phi}_k$'s in [10], [12]. This requires larger groups comprised of Mm^2 variables in the group penalty term $P_g(\mathbf{\Omega})$ given by (19), compared to groups of M variables in [10], [12]. In [10], [12] the group penalty term is missing the factor $\sqrt{M}$ (see [10, (41)]) corresponding to the factor $m\sqrt{M}$ in (19) of this paper. This factor equals the square-root of the number of group variables, following the work of [43]. A consequence of the extra factor $m\sqrt{M}$ is that in Theorem 1 of this paper, the bounds on λ_n for the lasso penalty do not depend on α (see (46) and (47)), whereas the corresponding result (with $m = 1$) in [10, (69)] depends upon α. In [10, Theorem 1] the lower bound on λ_n can be greater than the upperbound for certain choices of a constant C_1 whereas no such anomaly arises in this paper.

ii) In [10] non-convex penalties are not considered. In [12] non-convex log-sum regularization for CIG learning for single-attribute Gaussian time series has been proposed replacing the lasso penalty of [10]. It is shown empirically in [12] that the log-sum penalty significantly outperforms the lasso penalty with F_1 score as a performance measure. A theorem corresponding to Theorem 1 of this paper and that of [10] is stated in [12] without any proof. Moreover, as in [10], for lack of the factor $\sqrt{M}$ in the group penalty term in [12], the upperbound on λ_n in [12, Theorem] depends on α and it can be smaller than the lowerbound for certain choices of a constant. No such anomaly arises in this paper.

iii) In this paper we provide results for two non-convex penalties (SCAD and log-sum) for CIG learning from multi-attribute time series. The two penalties require different analysis in proving Theorem 1 (compare (85)–(95) for lasso and log-sum penalties with (97)–(102) for the SCAD penalty in Appendix A). We provide a complete proof of Theorem 1 whereas [12] has no proof of its theorem.

iv) There are no results corresponding to our Lemma 1 and Theorems 2 and 3 in [10], [12] (not needed in [10] since it does not consider non-convex penalties). □

VI. NUMERICAL EXAMPLES

In this section we present numerical results using both synthetic and real data to illustrate the proposed approach. We know the ground truth in the synthetic data example which permits assessment of the efficacy of our approaches. The ground truth is unknown in the real data example and here we wish to visualize and explore the conditional dependency structure underlying the data.

A. SYNTHETIC DATA

Consider a graph with $p = 64$ nodes, each node with $m = 4$ attributes. The time series data $\{\boldsymbol{x}(t)\}$ is generated using a vector autoregressive model of order 3 (VAR(3)):

$$\boldsymbol{x}(t) = \sum_{i=1}^{3} \boldsymbol{A}_i \boldsymbol{x}(t-i) + \boldsymbol{w}(t), \quad \boldsymbol{x}(t) \in \mathbb{R}^{mp}, \tag{60}$$

where $\boldsymbol{w}(t)$ is i.i.d. zero-mean Gaussian with precision matrix either $\tilde{\boldsymbol{\Omega}} = \tilde{\boldsymbol{\Omega}}_1$ (labeled Model 1) or $\tilde{\boldsymbol{\Omega}} = \tilde{\boldsymbol{\Omega}}_1 + \tilde{\boldsymbol{\Omega}}_2$ (labeled Model 2). For Model 1, we create 8 clusters (communities) of 8 nodes each, each node with $m = 4$ attributes, where nodes within a community are not connected to any node in other communities. To generate $\tilde{\boldsymbol{\Omega}}_1$, we set $[\tilde{\boldsymbol{\Omega}}_1^{(q\ell)}]_{uv} = 0.5^{|u-v|}$ for $q = \ell \in [8], u \neq v,\ u, v \in [m]$ (notation as in 2), and it is zero otherwise. For $q \neq \ell$, we have $\tilde{\boldsymbol{\Omega}}_1^{(q\ell)} = \boldsymbol{0}$. We add $\gamma \boldsymbol{I}_{mp}$ to $\tilde{\boldsymbol{\Omega}}_1$ and choose γ to make the minimum eigenvalue of $\tilde{\boldsymbol{\Omega}}_1 + \gamma \boldsymbol{I}_{mp}$ equal to 0.5 . The parameters of VAR(3) model are generated similarly by having $\boldsymbol{A}_i^{(q\ell)} = \boldsymbol{0}$ for $q \neq \ell$, and only 10% of the entries of $\boldsymbol{A}_i^{(qq)}$'s are nonzero with the nonzero elements independently and uniformly distributed over $[-0.6, 0.6]$. We then check if the VAR(3) model is stable, a necessary and sufficient condition for which is that the roots of $a(z) = |\boldsymbol{I}_{mp} - \sum_{i=1}^{3} \boldsymbol{A}_i z^{-i}| = 0$ should all have modulus < 1; this condition is equivalent to having all eigenvalues of the corresponding $(3mp) \times (3mp)$ companion matrix to have modulus < 1 [47, Sec. 8.2.3]. Additionally, in order to avoid a "long" impulse response, we require the roots of $a(z)$ to have modulus ≤ 0.95. Suppose this condition is violated with $|z_{\max}| > 0.95$ where $|z_{\max}| = \arg\max_{\ell \in [3mp]} \{|z_\ell| \,:\, a(z_\ell) = 0\}$. In this case, we scale $\boldsymbol{A}_i$'s to $\bar{\boldsymbol{A}}_i = \gamma^i \boldsymbol{A}_i$, $\gamma = 0.95/|z_{\max}|$. It is easy to see that the roots of $\bar{a}(z) = |\boldsymbol{I}_{mp} - \sum_{i=1}^{3} \bar{\boldsymbol{A}}_i z^{-i}| = a(z/\gamma) = 0$ now all have modulus ≤ 0.95.

For Model 2, we allow some interaction between the 8 clusters via $\tilde{\boldsymbol{\Omega}}_2$ which is generated via an Erdös-Rènyi graph structure where the p nodes are connected with probability $p_{er} = 0.002$. To generate $\tilde{\boldsymbol{\Omega}}_2$, we set $\tilde{\boldsymbol{\Omega}}_2^{(q\ell)} = \boldsymbol{0}$ for $q = \ell \in [8]$, and for $q \neq \ell$ but connected in the Erdös-Rènyi graph, the entries of $\tilde{\boldsymbol{\Omega}}_2^{(q\ell)}$ are independently and uniformly distributed over $[-0.4, -0.1] \cup [0.1, 0.4]$, and are zero if not connected.

First 100 samples are discarded to eliminate transients. This set-up leads to approximately 11% and 13% connected edges in models 1 and 2, respectively. In each run, we calculated the true PSD $\boldsymbol{S}(f)$ for $f \in [0, 0.5]$ at intervals of 0.01, and then take $\{q, \ell\} \in \mathcal{E}$ if $\sqrt{\sum_f \|(\boldsymbol{S}^{-1}(f))^{(q\ell)}\|_F^2} > 10^{-2} \left(\max_{q,\ell \in [p]} \sqrt{\sum_f \|(\boldsymbol{S}^{-1}(f))^{(q\ell)}\|_F^2}\right)$, else $\{q, \ell\} \notin \mathcal{E}$. For a typical realization (run), Fig. 1 shows heatmaps of $\log_{10}\left(\sum_{f=0:0.01:5} |[S^{-1}(f)]_{ij}|\right)$, $i, j \in [256]$, for models 1 and 2.

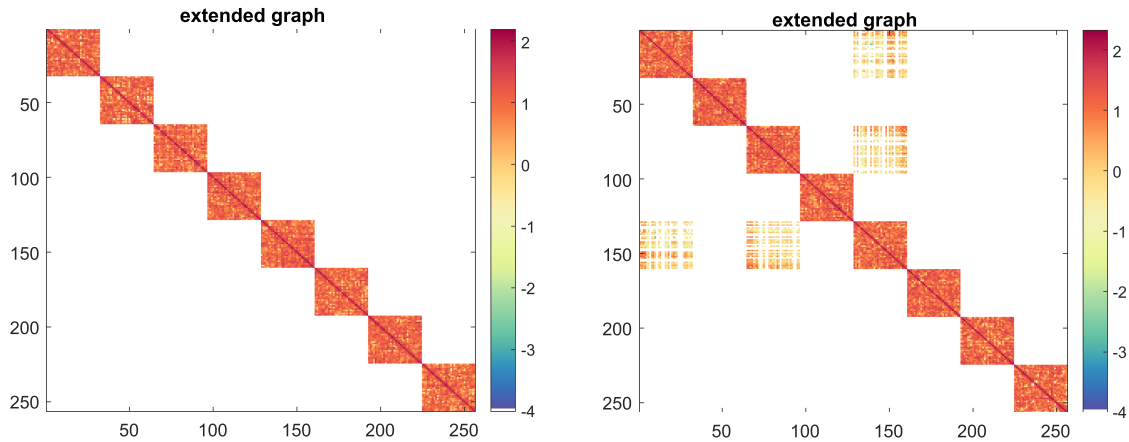
Simulation results based on 100 runs are shown in Tables 1–4 where the performance measures are F_1-score and Hamming distance for efficacy in edge detection. All

TABLE 1. Model 1: F_1 Scores, Hamming Distances and Timings, Averaged Over 100 Runs

n	128	256	1024
M=4: F_1 score $\pm\sigma$: λ's picked to maximize F_1			
Lasso	0.5788 ± 0.1407	0.7647 ± 0.1308	0.9682 ± 0.0347
Log-sum	0.7065 ± 0.0517	0.8679 ± 0.0261	0.9899 ± 0.0077
SCAD	0.5820 ± 0.1428	0.7651 ± 0.1312	0.9675 ± 0.0347
M=4: Hamming distance $\pm\sigma$: λ's picked to maximize F_1			
Lasso	168.53 ± 040.255	097.36 ± 044.03	013.93 ± 014.71
Log-sum	113.32 ± 012.37	057.70 ± 011.05	004.46 ± 003.34
SCAD	165.41 ± 037.59	097.14 ± 044.15	014.19 ± 014.66
M=4: Timing (s) $\pm\sigma$: λ's picked to maximize F_1			
Lasso	011.45 ± 01.105	009.52 ± 01.477	005.65 ± 00.585
Log-sum	019.62 ± 00.497	016.88 ± 01.309	010.94 ± 00.808
SCAD	023.79 ± 02.005	019.29 ± 02.774	011.92 ± 01.206
M=4: F_1 score $\pm\sigma$: λ's picked to minimize BIC			
Log-sum	0.4394 ± 0.0106	0.6632 ± 0.0496	0.9577 ± 0.0534
M=4: Hamming distance $\pm\sigma$: λ's picked to minimize BIC			
Log-sum	499.97 ± 015.93	214.10 ± 050.73	017.22 ± 020.15

TABLE 2. Model 2: F_1 Scores, Hamming Distances and Timings, Averaged Over 100 Runs

n	128	256	1024
M=4: F_1 score $\pm\sigma$: λ's picked to maximize F_1			
Lasso	0.4907 ± 0.0853	0.6098 ± 0.1460	0.7847 ± 0.0911
Log-sum	0.5692 ± 0.0346	0.7241 ± 0.0599	0.8236 ± 0.0696
SCAD	0.4982 ± 0.0862	0.6087 ± 0.1456	0.7819 ± 0.0911
M=4: Hamming distance $\pm\sigma$: λ's picked to maximize F_1			
Lasso	307.37 ± 115.32	219.54 ± 147.61	126.68 ± 103.50
Log-sum	241.40 ± 038.08	145.72 ± 043.82	099.39 ± 047.80
SCAD	387.51 ± 115.59	219.77 ± 147.05	127.69 ± 102.12
M=4: Timing (s) $\pm\sigma$: λ's picked to maximize F_1			
Lasso	10.598 ± 01.167	09.090 ± 01.534	06.790 ± 01.190
Log-sum	20.141 ± 00.719	16.882 ± 01.519	12.783 ± 01.526
SCAD	20.555 ± 02.474	17.446 ± 02.908	12.562 ± 01.831
M=4: F_1 score $\pm\sigma$: λ's picked to minimize BIC			
Log-sum	0.5136 ± 0.0195	0.7224 ± 0.0644	0.7623 ± 0.1030
M=4: Hamming distance $\pm\sigma$: λ's picked to minimize BIC			
Log-sum	359.59 ± 039.91	139.84 ± 039.46	115.90 ± 050.71



(a) Model 1 (b) Model 2

FIGURE 1. True $\log_{10}(\sum_{f=0:0.01:5} |[S^{-1}(f)]_{ij}|)$, $i, j \in [256]$, for extended graphs for a single Monte Carlo run: $mp = 4 \times 64 = 256$ nodes.

TABLE 3. Model 2: F_1 Scores and Hamming Distances Using Log-Sum Penalty, Averaged Over 100 Runs

n	128	256	1024
F_1 score $\pm\sigma$: λ's picked to maximize F_1			
M=2	0.5967 ± 0.0440	0.7324 ± 0.0692	0.8283 ± 0.0695
M=3	0.5826 ± 0.0396	0.7294 ± 0.0664	0.8269 ± 0.0691
M=4	0.5692 ± 0.0346	0.7241 ± 0.0599	0.8236 ± 0.0696
M=6	0.5156 ± 0.0358	0.7046 ± 0.0611	0.8158 ± 0.0716
Hamming distance $\pm\sigma$: λ's picked to maximize F_1			
M=2	215.13 ± 039.42	134.40 ± 039.84	096.64 ± 047.14
M=3	224.72 ± 037.49	136.61 ± 040.49	097.85 ± 047.42
M=4	241.40 ± 038.08	145.72 ± 043.82	099.39 ± 047.80
M=6	261.21 ± 036.99	149.50 ± 038.84	104.11 ± 051.47

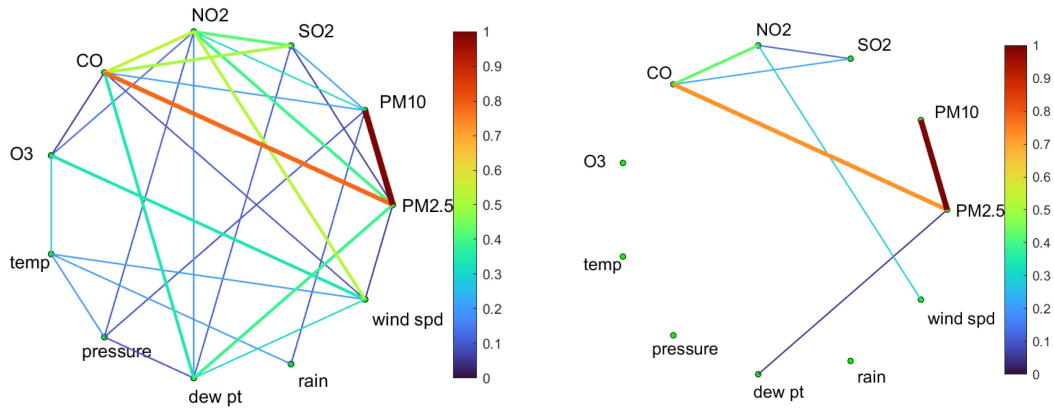
TABLE 4. Model 2, Varying AR Model Order: VAR(L) as in (61), $L \in \{1, 2, 3, 4\}$. F_1 Scores and Hamming Distances Using Log-Sum Penalty, Averaged Over 100 Runs

n	128	256	1024
M=4: F_1 score $\pm\sigma$: λ's picked to maximize F_1			
L=1	0.5045 ± 0.0273	0.6584 ± 0.0522	0.8067 ± 0.0500
L=2	0.5834 ± 0.0726	0.7050 ± 0.0622	0.8212 ± 0.0703
L=3	0.5692 ± 0.0346	0.7241 ± 0.0599	0.8236 ± 0.0696
L=4	0.5609 ± 0.0476	0.7191 ± 0.0625	0.8361 ± 0.0632
M=4: Hamming distance $\pm\sigma$: λ's picked to maximize F_1			
L=1	258.80 ± 028.79	161.73 ± 046.36	097.67 ± 032.57
L=2	202.80 ± 046.89	156.61 ± 052.88	098.76 ± 046.74
L=3	241.40 ± 038.08	145.72 ± 043.82	099.39 ± 047.80
L=4	248.40 ± 046.51	149.73 ± 046.00	090.37 ± 041.32

algorithms were run on a Window 10 Pro operating system with processor Intel(R) Core(TM) i7-10700 CPU @2.90GHz with 32GB RAM, using MATLAB R2023a. The F_1-score is defined as $F_1 = 2 \times \text{precision} \times \text{recall}/(\text{precision} + \text{recall})$ where $\text{precision} = |\hat{\mathcal{E}} \cap \mathcal{E}_0|/|\hat{\mathcal{E}}|$, $\text{recall} = |\hat{\mathcal{E}} \cap \mathcal{E}_0|/|\mathcal{E}_0|$, and $\mathcal{E}_0$ and $\hat{\mathcal{E}}$ denote the true and estimated edge sets, respectively. The Hamming distance is between $\hat{\mathcal{E}}$ and $\mathcal{E}_0$, scaled by 0.5 to count only distinct edges. For our proposed approach, we consider $M = 4$ for three samples sizes $n \in \{128, 256, 1024\}$ in Table 1 for Model 1 and Table 2 for Model 2. For $M = 4$, we used $K = 2m_t + 1 = 15, 31, 127$ for $n = 128, 256, 1024$, respectively. We fixed $\alpha = 0.05$ and λ was selected by searching over a grid of values to maximize the F_1-score (over 100 runs), or via BIC as in Section IV-A. We used lasso (convex), log-sum (non-convex, $\epsilon = 0.0001$) or SCAD (non-convex, a=3.7) penalties. When λ's are picked to maximize the F_1 score, it is seen that the log-sum penalty outperforms the lasso and the SCAD penalties in both Table 1 (Model 1) and Table 2 (Model 2) in terms of the F_1-score as well as the Hamming distance, whereas the SCAD penalty does not offer much improvement over lasso. As discussed in Remark 2, the "convexity" region for the log-sum penalty is likely to be much larger than that for SCAD. With the lasso penalty, computational time is close to half of that for log-sum or SCAD, which is not surprising since the latter are initialized using the lasso result (cf. Section IV). When λ's are picked via BIC (only for the log-sum penalty), there is a drop in the F_1 score and increase in the Hamming distance as compared to the case where λ's are picked to maximize the F_1 score. This is due to errors in the BIC parameter selection method.

In Table 3 we show the results for the log-sum penalty for $M = 2, 3, 4$ and 6 with λ selected to maximize the F_1-score. We take $n = 128, 256, 1024$ and the corresponding m_t values leading to different M values are $m_t = 15, 31, 127$ ($M = 2$), $m_t = 9, 20, 84$ ($M = 3$), $m_t = 7, 15, 63$ ($M = 4$), and $m_t = 4, 10, 42$ ($M = 6$). The number of unknown parameters being estimated are $\mathcal{O}(M(mp)^2)$ for M $(mp) \times (mp)$ $\mathbf{\Phi}_k$'s. We see that for a fixed n, at first the performance changes only a little with increasing M, then it declines more sharply ($M = 4$ to 6) as more parameters need to be estimated with increasing M.

In Table 4 we display some numerical ablation results by varying the AR model order. We use a VAR(L) model with



(a) M=4: Lasso penalty

(b) M=4: Log-sum penalty

FIGURE 2. Pollution graphs for the Beijing air-quality dataset [24] for year 2013-14: 8 monitoring sites and 11 features ($m = 8$, $p = 11$, $M = 4$, $n = 364$). Number of distinct edges = 29 and 7 in graphs (a) and (b), respectively. Estimated $\|\hat{\Omega}^{(ijM)}\|_F$ is the edge weight (normalized to have $\max_{i\neq j} \|\hat{\Omega}^{(ijM)}\|_F = 1$), see (20). The edge weights are color coded, in addition to the edges with higher weights being drawn thicker.

$L \in \{1, 2, 3, 4\}$, given by

$$\boldsymbol{x}(t) = \sum_{i=1}^{L} \boldsymbol{A}_i \boldsymbol{x}(t-i) + \boldsymbol{w}(t), \quad \boldsymbol{x}(t) \in \mathbb{R}^{mp}, \tag{61}$$

where $\boldsymbol{A}_i$'s are picked as for (60) and we used Model 2 to specify the precision matrix of $\boldsymbol{w}(t)$. We used the log-sum penalty and $M = 4$. It is seen that the results are consistent across model orders, both in terms of the F_1 scores and the Hamming distances.

B. REAL DATA: BEIJING AIR-QUALITY DATASET [24]

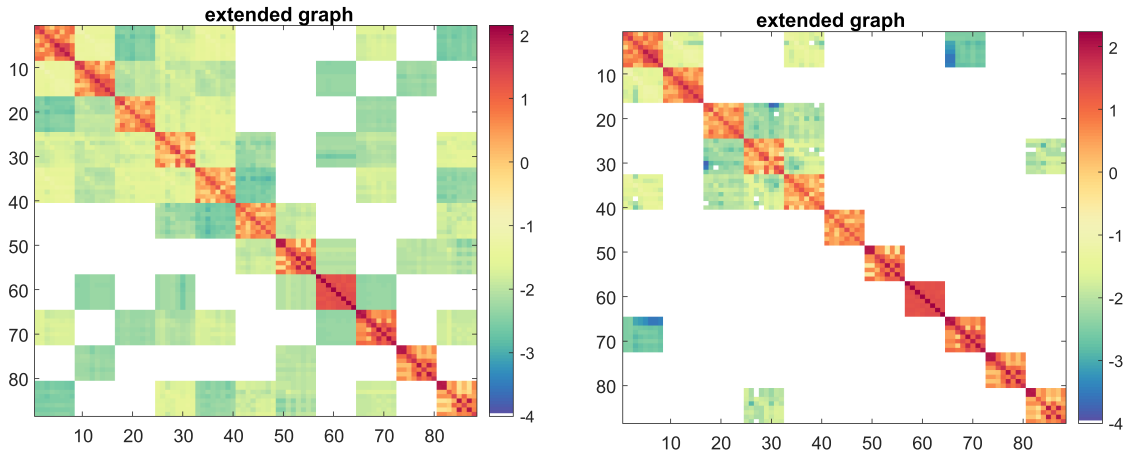
Here we consider Beijing air-quality dataset [24], [25], downloaded from https://archive.ics.uci.edu/dataset/501/beijing+multi+site+air+quality+data. This data set includes hourly air pollutants data from 12 nationally-controlled air-quality monitoring sites in the Beijing area. The time period is from March 1st, 2013 to February 28th, 2017. The six air pollutants are $PM_{2.5}$, PM_{10}, SO_2, NO_2, CO, and O_3, and the meteorological data is comprised of five features: temperature, atmospheric pressure, dew point, wind speed, and rain; we did not use wind direction. Thus we have eleven ($= p$) features (pollutants and weather variables). We used data from 8 ($= m$) sites: Changping, Dingling, Huairou, Shunyi, Aotizhongxin, Dongsi, Guanyuan, Gucheng. The data are averaged over 24 h period to yield daily averages $x_i(t)$, $i \in [88]$. We used one year 2013-14 of daily data resulting in $n = 365$ days. We pre-processed the data as follows. Given $x_i(t)$, we transform it to $\bar{x}_i(t) = \ln(x_i(t)/x_i(t-1))$ for each i (leads to $n = 364$), and then detrend it (i.e., remove the best straight-line fit). Finally, we scale the detrended scalar sequence to have a mean-square value of one. All temperatures were converted from Celsius to Kelvin to avoid negative numbers. If a value of a feature is zero (e.g., wind speed), we added a small positive number to it so that the log transformation is well-defined.

Fig. 2 shows the CIGs for lasso and log-sum penalties for $M = 4$ where with $\alpha = 0.05$, λ was selected via BIC: an edges exists iff $\|\hat{\boldsymbol{\Omega}}^{(ijM)}\|_F > 0$, $i, j \in [11]$. The corresponding heatmaps showing estimated $\log_{10}(\sqrt{\sum_{k=1}^{M} |[\hat{\boldsymbol{\Phi}}_k]_{ij}|^2})$, $i, j \in [88]$ are in Fig. 3. It is seen that lasso yields a much denser graph (29 edges) while the graph resulting from the log-sum penalty is much sparser (7 edges). Cold, dry air from the north of Beijing reduces both dew point and $PM_{2.5}$ particle concentration in suburban areas while southerly wind brings warmer and more humid air from the more polluted south that elevates both dew point and $PM_{2.5}$ concentration [24]. This fact is captured by the edge between dew point and $PM_{2.5}$ in Fig. 2.

VII. CONCLUSION

Estimation of the CIG of high-dimensional multivariate Gaussian time series from multi-attribute data was considered. We provided a unified theoretical analysis of multi-attribute graph learning for dependent time series using a penalized log-likelihood objective function in the frequency-domain. Both convex and non-convex regularization functions were considered. We established sufficient conditions for consistency, local convexity when using non-convex penalties, and graph recovery. Our approach was illustrated using numerical examples utilizing both synthetic and real (Beijing air-quality dataset) data. Non-convex log-sum regularization yielded more accurate results compared to convex sparse-group lasso or non-convex SCAD regularization for synthetic data, and sparser graph for real data.

Now we briefly outline several avenues for future work in this area. Although we observe empirically that the log-sum penalty significantly outperforms the lasso penalty, we



(a) Lasso (b) Log-sum

FIGURE 3. Estimated $\log_{10}(\sqrt{\sum_{k=1}^{M} |[\hat{\Phi}_k]_{ij}|^2})$, $i, j \in [88]$, for the Beijing air-quality dataset ($m = 8$, $p = 11$, $M = 4$, $n = 364$). There are $p = 11$ nodes ($PM_{2.5}$ labeled as node 1, PM_{10} as 2, and so on, moving counter-clockwise in Fig. 2), each variables measured at $m = 8$ stations.

have not proved it. This would be a task for future research. Theorem 1 shows that all three penalties considered yield the same asymptotic rate of convergence (see Remark 1) and how this rate is influenced by various chosen parameters. Other model selection approaches also need to be investigated. In [51] a stability approach is proposed for graphical modeling with i.i.d. data using a random sampling approach to pick the smallest regularization parameter that "simultaneously makes the graph sparse and replicable under random sampling." In our case, in the frequency-domain we have M models, each with K approximately i.i.d. complex-valued Gaussian measurements $\boldsymbol{d}_x(\tilde{f}_{k,\ell})$ (see Section II–II-A). In order to apply the approach of [51], we would sample in the frequency-domain. Finally, sample complexity issues based on information-theoretic bounds need to be investigated. Based on the results of [52] for i.i.d. data, [53] consider stationary Gaussian sequences, and using information-theoretic methods, [53] derive a lower bound on the error probability of any learning scheme for the underlying process CIG. This bound is then used to derive a minimum required sample-size which is necessary for any algorithm regardless of its computational complexity, to reliably select the true underlying CIG. The model restrictions in [53] are more stringent than we consider here; e.g., instead of our assumption (A1), [53] needs $\sum_{\tau=-\infty}^{\infty} |\tau| \, |[\boldsymbol{R}_{xx}(\tau)]_{k\ell}| < \infty$ for every $k, \ell \in \bar{V}$.

APPENDIX

A. PROOF OF THEOREM 1

Our proof relies on the method of [48] which deals with i.i.d. time series models and lasso penalty, and our prior results in [10] dealing with sparse group lasso penalty and single-attribute time series. From now on we use the term "with high probability" (w.h.p.) to denote with probability greater than $1 - 1/(mp_n)^{\tau-2}$. First we recall Lemmas 2 and 3 from [10], replacing p_n therein with mp_n.

We denote $\boldsymbol{S}_0(\tilde{f}_k)$ as $\boldsymbol{S}_{0k}$ in this section.

Lemma 2: [10, Lemma 3]. Under Assumption (A1)-(A2), $\hat{\boldsymbol{S}}_k$ satisfies the tail bound

$$P\left(\max_{k,q,l} |[\hat{\boldsymbol{S}}_k - \boldsymbol{S}_{0k}]_{ql}| > C_0 \sqrt{\frac{\ln(mp_n)}{K_n}}\right) \leq \frac{1}{(mp_n)^{\tau-2}}$$

for $\tau > 2$, if the sample size $n > N_1$, where C_0 is defined in (38) and N_1 is defined in (42). •

Lemma 3 deals with a Taylor series expansion with integral remainder using Wirtinger calculus [49].

Lemma 3: [10, Lemma 5]. With $c(\boldsymbol{\Phi}_k, \boldsymbol{\Phi}_k^*) = \ln|\boldsymbol{\Phi}_k| + \ln|\boldsymbol{\Phi}_k^*|$ and $\boldsymbol{\Phi}_k = \boldsymbol{\Phi}_{0k} + \boldsymbol{\Gamma}_k = \boldsymbol{\Phi}_k^H$, the Taylor series expansion of $c(\boldsymbol{\Phi}_k, \boldsymbol{\Phi}_k^*)$ in the integral remainder form is given by

$$\begin{aligned} c(\boldsymbol{\Phi}_k, \boldsymbol{\Phi}_k^*) &= c(\boldsymbol{\Phi}_{0k}, \boldsymbol{\Phi}_{0k}^*) + \mathrm{tr}(\boldsymbol{\Phi}_{0k}^{-1}\boldsymbol{\Gamma}_k + \boldsymbol{\Phi}_{0k}^{-*}\boldsymbol{\Gamma}_k^*) \\ &\quad - \boldsymbol{g}^H(\boldsymbol{\Gamma}_k)\left(\int_0^1 (1-v)\boldsymbol{H}(\boldsymbol{\Phi}_{0k}, \boldsymbol{\Gamma}_k, v)\, dv\right)\boldsymbol{g}(\boldsymbol{\Gamma}_k) \end{aligned} \tag{62}$$

where v is real,

$$\boldsymbol{g}(\boldsymbol{\Gamma}_k) = \begin{bmatrix} \mathrm{vec}(\boldsymbol{\Gamma}_k) \\ \mathrm{vec}(\boldsymbol{\Gamma}_k^*) \end{bmatrix}, \quad \boldsymbol{H}(\boldsymbol{\Phi}_{0k}, \boldsymbol{\Gamma}_k, v) = \begin{bmatrix} \boldsymbol{H}_{11k} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{H}_{22k} \end{bmatrix} \tag{63}$$

$$\boldsymbol{H}_{11k} = (\boldsymbol{\Phi}_{0k} + v\boldsymbol{\Gamma}_k)^{-*} \otimes (\boldsymbol{\Phi}_{0k} + v\boldsymbol{\Gamma}_k)^{-1}, \tag{64}$$

and

$$\boldsymbol{H}_{22k} = (\boldsymbol{\Phi}_{0k} + v\boldsymbol{\Gamma}_k)^{-1} \otimes (\boldsymbol{\Phi}_{0k} + v\boldsymbol{\Gamma}_k)^{-*} \quad \bullet \tag{65}$$

We now turn to the proof of Theorem 1.

Proof of Theorem 1: Let $\boldsymbol{\Omega} = \boldsymbol{\Omega}_0 + \boldsymbol{\Delta}$ where

$$\boldsymbol{\Delta} = \begin{bmatrix} \boldsymbol{\Gamma}_1 & \boldsymbol{\Gamma}_2 & \cdots & \boldsymbol{\Gamma}_{M_n} \end{bmatrix} \in \mathbb{C}^{(mp_n)\times(mp_nM_n)}, \tag{66}$$

$$\mathbf{\Gamma}_k = \mathbf{\Phi}_k - \mathbf{\Phi}_{0k} \in \mathbb{C}^{(mp_n)\times(mp_n)},\ k \in [M_n], \tag{67}$$

and $\mathbf{\Phi}_k$, $\mathbf{\Phi}_{0k}$ are both Hermitian positive-definite, implying $\mathbf{\Gamma}_k = \mathbf{\Gamma}_k^H$. Let

$$Q(\mathbf{\Omega}) := \bar{\mathcal{L}}(\mathbf{\Omega}) - \bar{\mathcal{L}}(\mathbf{\Omega}_0)\,. \tag{68}$$

The estimate $\hat{\mathbf{\Omega}}_\lambda$, denoted by $\hat{\mathbf{\Omega}}$ hereafter suppressing dependence upon λ, minimizes $Q(\mathbf{\Omega})$, or equivalently, $\hat{\mathbf{\Delta}} = \hat{\mathbf{\Omega}} - \mathbf{\Omega}_0$ minimizes $G(\mathbf{\Delta}) := Q(\mathbf{\Omega}_0 + \mathbf{\Delta})$. We will follow the proof of [10, Theorem 1], which, in turn, follows the method of proof of [48, Theorem 1] pertaining to real-valued i.i.d. time series. Consider the set

$$\Theta_n(R) := \left\{\mathbf{\Delta} \,:\, \mathbf{\Gamma}_k = \mathbf{\Gamma}_k^H,\ k \in [M_n],\ \|\mathbf{\Delta}\|_F = Rr_n\right\} \tag{69}$$

where R and r_n are as in (40) and (41), respectively. Observe that

$$G(\hat{\mathbf{\Delta}}) = Q(\mathbf{\Omega}_0 + \hat{\mathbf{\Delta}}) \le G(\mathbf{0}) = 0\,. \tag{70}$$

Therefore, if we can show that

$$\inf_{\mathbf{\Delta}}\{G(\mathbf{\Delta}) \,:\, \mathbf{\Delta} \in \Theta_n(R)\} \,>\, 0\,, \tag{71}$$

a minimizer $\hat{\mathbf{\Delta}}$ must be inside the sphere defined by $\Theta_n(R)$, and hence

$$\|\hat{\mathbf{\Delta}}\|_F \le Rr_n\,. \tag{72}$$

When $G(\mathbf{\Delta})$ is convex (as with the lasso penalty), the minimizer is global, else it is local.

Using Lemma 3 we rewrite $G(\mathbf{\Delta})$ as

$$G(\mathbf{\Delta}) = \sum_{k=1}^{M_n}\left(\frac{1}{2}A_{1k} + \frac{1}{2}A_{2k} + A_{3k}\right) + A_4\,, \tag{73}$$

where, noting that $\mathbf{\Phi}_{0k}^{-1} = \boldsymbol{S}_{0k}$,

$$A_{1k} = \boldsymbol{g}^H(\mathbf{\Gamma}_k)\left(\int_0^1 (1-v)\boldsymbol{H}(\mathbf{\Phi}_{0k}, \mathbf{\Gamma}_k, v)\,dv\right)\boldsymbol{g}(\mathbf{\Gamma}_k)\,, \tag{74}$$

$$A_{2k} = \mathrm{tr}\left((\hat{\boldsymbol{S}}_k - \boldsymbol{S}_{0k})\mathbf{\Gamma}_k + (\hat{\boldsymbol{S}}_k - \boldsymbol{S}_{0k})^*\mathbf{\Gamma}_k^*\right)\,, \tag{75}$$

$$A_{3k} = \alpha\sum_{i\ne j}^{mp_n}\left(\rho_\lambda\left([\mathbf{\Phi}_{0k} + \mathbf{\Gamma}_k]_{ij}\right) - \rho_\lambda\left([\mathbf{\Phi}_{0k}]_{ij}\right)\right)\,, \tag{76}$$

$$A_4 = (1-\alpha)m\sqrt{M}\sum_{q\ne\ell}^{p_n}\left(\rho_\lambda\left(\|\mathbf{\Omega}_0^{(q\ell M_n)} + \mathbf{\Delta}^{(q\ell M_n)}\|_F\right) - \rho_\lambda\left(\|\mathbf{\Omega}_0^{(q\ell M_n)}\|_F\right)\right)\,, \tag{77}$$

$$\mathbf{\Omega}_0^{(q\ell M_n)} := \left[\mathbf{\Phi}_1^{(q\ell)},\ \mathbf{\Phi}_2^{(q\ell)},\ \cdots,\ \mathbf{\Phi}_{M_n}^{(q\ell)}\right] \in \mathbb{C}^{m\times(mM_n)}\,, \tag{78}$$

$$\mathbf{\Delta}^{(q\ell M_n)} := \left[\mathbf{\Gamma}_1^{(q\ell)},\ \mathbf{\Gamma}_2^{(q\ell)},\ \cdots,\ \mathbf{\Gamma}_{M_n}^{(q\ell)}\right] \in \mathbb{C}^{m\times(mM_n)}\,. \tag{79}$$

Also define

$$A_\ell = 0.5\sum_{k=1}^{M_n}A_{\ell k},\ \ \ell = 1, 2,\ \ A_3 = \sum_{k=1}^{M_n}A_{3k}\,, \tag{80}$$

and

$$d_{1n} := \sqrt{\frac{\ln(mp_n)}{K_n}},\ \ d_{2n} := d_{1n}\sqrt{mp_n + m^2 s_{n0}}\,. \tag{81}$$

The bounds on A_{1k}'s and A_1 follows exactly as in [10, Theorem 1], with the final result (see equations [10, (B.39)-(B.44)])

$$A_1 \ge \frac{\|\mathbf{\Delta}\|_F^2}{2\left(\beta_{\min}^{-1} + Rr_n\right)^2}\,. \tag{82}$$

Turning to A_{2k}'s and A_2, as in [10, Theorem 1] (after accounting for the fact that here we have $\hat{\boldsymbol{S}}_k \in \mathbb{C}^{(mp_n)\times(mp_n)}$ whereas in [10], $\hat{\boldsymbol{S}}_k \in \mathbb{C}^{p_n\times p_n}$, and here we have group penalty on groups of size m^2M_n elements whereas in [10], group size is M_n), with probability $>\ 1 - 1/(mp_n)^{\tau-2}$, we have the bound [10, (B.51)]

$$|A_2| \le C_0\sum_{k=1}^{M_n}\left(d_{1n}\|\mathbf{\Gamma}_k^-\|_1 + d_{2n}\|\mathbf{\Gamma}_k^+\|_F\right) \tag{83}$$

a well as the bound [10, (B.55)]

$$|A_2| \le \sqrt{m^2M_n}\,C_0\,d_{1n}\left(\|\tilde{\mathbf{\Delta}}^-\|_1 + \|\tilde{\mathbf{\Delta}}^+\|_1\right) \tag{84}$$

where $\tilde{\mathbf{\Delta}} \in \mathbb{R}^{p_n\times p_n}$ has its (i, j)th element $\tilde{\Delta}_{ij} = \|\mathbf{\Delta}^{(ijM_n)}\|_F$.

For the rest of the proof we have two slightly different approaches, one for lasso and log-sum and the other for SCAD penalty. The following applies to lasso and log-sum penalties.

For Lasso and Log-Sum Penalties: We now bound A_{3k}. Let $\mathcal{E}_0^c$ denote the complement of $\mathcal{E}_0$, given by $\mathcal{E}_0^c = \{\{i, j\} : (\boldsymbol{S}_0^{-1}(f))^{(ij)} \equiv \mathbf{0},\ i \ne j,\ i, j \in [p_n],\ f \in [0, 0.5]\}$. Similarly, let $\bar{\mathcal{E}}_0^c$ denote the complement of $\bar{\mathcal{E}}_0$, given by $\bar{\mathcal{E}}_0^c = \{\{i, j\} : [\boldsymbol{S}_0^{-1}(f)]_{ij} \equiv 0,\ i \ne j,\ i, j \in [mp_n],\ f \in [0, 0.5]\}$. For an index set $\boldsymbol{B}$ and a matrix $\boldsymbol{C} \in \mathbb{C}^{q\times q}$, we write $\boldsymbol{C}_{\boldsymbol{B}}$ to denote a matrix in $\mathbb{C}^{q\times q}$ such that $[\boldsymbol{C}_{\boldsymbol{B}}]_{ij} = C_{ij}$ if $(i, j) \in \boldsymbol{B}$, and $[\boldsymbol{C}_{\boldsymbol{B}}]_{ij} = 0$ if $(i, j) \notin \boldsymbol{B}$. Then $\mathbf{\Gamma}_k^- = \mathbf{\Gamma}_{k\mathcal{E}_0}^- + \mathbf{\Gamma}_{k\mathcal{E}_0^c}^-$, and $\|\mathbf{\Gamma}_k^-\|_1 = \|\mathbf{\Gamma}_{k\mathcal{E}_0}^-\|_1 + \|\mathbf{\Gamma}_{k\mathcal{E}_0^c}^-\|_1$. We have

$$\begin{aligned} A_{3k} &= \alpha\sum_{(i,j)\in\bar{\mathcal{E}}_0}\left(\rho_\lambda\left([\mathbf{\Phi}_{0k} + \mathbf{\Gamma}_k]_{ij}\right) - \rho_\lambda\left([\mathbf{\Phi}_{0k}]_{ij}\right)\right) \\ &\quad + \alpha\sum_{(i,j)\in\bar{\mathcal{E}}_0^c}\rho_\lambda\left([\mathbf{\Gamma}_k]_{ij}\right) \\ &= \alpha\sum_{(i,j)\in\bar{\mathcal{E}}_0}\rho_\lambda'\left([\tilde{\mathbf{\Phi}}_k]_{ij}\right)\left(|[\mathbf{\Phi}_{0k} + \mathbf{\Gamma}_k]_{ij}| - |[\mathbf{\Phi}_{0k}]_{ij}|\right) \\ &\quad + \alpha\sum_{(i,j)\in\bar{\mathcal{E}}_0^c}\rho_\lambda\left([\mathbf{\Gamma}_k]_{ij}\right) \end{aligned} \tag{85}$$

where we used the mean value theorem to infer $\rho_\lambda([\mathbf{\Phi}_{0k} + \mathbf{\Gamma}_k]_{ij}) = \rho_\lambda([\mathbf{\Phi}_{0k}]_{ij}) + \rho_\lambda'([\tilde{\mathbf{\Phi}}_k]_{ij})(|[\mathbf{\Phi}_{0k} + \mathbf{\Gamma}_k]_{ij}| - |[\mathbf{\Phi}_{0k}]_{ij}|)$ for some $|[\tilde{\mathbf{\Phi}}_k]_{ij}| = |[\mathbf{\Phi}_{0k}]_{ij}| + \gamma(|[\mathbf{\Phi}_{0k} + \mathbf{\Gamma}_k]_{ij}| - |[\mathbf{\Phi}_{0k}]_{ij}|)$ and $\gamma \in [0, 1]$. Using the triangle inequality, properties (vii) and (viii) of the penalty functions,

and $C_\lambda = \lambda/2$, we have

$$A_{3k} \geq -\alpha \sum_{(i,j)\in\bar{\mathcal{E}}_0} \rho'_\lambda\left([\tilde{\mathbf{\Phi}}_k]_{ij}\right) |[\mathbf{\Gamma}_k]_{ij}| + \alpha \sum_{(i,j)\in\bar{\mathcal{E}}_0^c} C_\lambda |[\mathbf{\Gamma}_k]_{ij}| \text{ for } |[\mathbf{\Gamma}_k]_{ij}| \leq \delta_\lambda \quad (86)$$

$$\geq -\alpha\lambda_n \sum_{(i,j)\in\bar{\mathcal{E}}_0} |[\mathbf{\Gamma}_k]_{ij}| + \frac{\alpha\lambda_n}{2} \sum_{(i,j)\in\bar{\mathcal{E}}_0^c} |[\mathbf{\Gamma}_k]_{ij}|$$

$$= \alpha\lambda_n \left(\frac{1}{2}\|\mathbf{\Gamma}^-_{k\bar{\mathcal{E}}_0^c}\|_1 - \|\mathbf{\Gamma}^-_{k\bar{\mathcal{E}}_0}\|_1\right), \quad (87)$$

leading to ($A_3 = \sum_{k=1}^{M_n} A_{3k}$)

$$A_3 \geq \alpha\lambda_n \sum_{k=1}^{M_n} \left(\frac{1}{2}\|\mathbf{\Gamma}^-_{k\bar{\mathcal{E}}_0^c}\|_1 - \|\mathbf{\Gamma}^-_{k\bar{\mathcal{E}}_0}\|_1\right). \quad (88)$$

Similarly, by (77), we have

$$A_4 \geq (1-\alpha)m\sqrt{M_n}\lambda_n \left(\frac{1}{2}\sum_{(q,\ell)\in\mathcal{E}_0^c} \|\mathbf{\Delta}^{(q\ell M_n)}\|_F - \sum_{(q,\ell)\in\mathcal{E}_0} \|\mathbf{\Delta}^{(q\ell M_n)}\|_F\right). \quad (89)$$

Now $\|\mathbf{\Gamma}^-_{k\bar{\mathcal{E}}_0}\|_1 \leq \sqrt{m^2 s_{n0}}\,\|\mathbf{\Gamma}^-_{k\bar{\mathcal{E}}_0}\|_F \leq \sqrt{m^2 s_{n0}}\,\|\mathbf{\Gamma}_k\|_F$, by the Cauchy-Schwarz inequality, hence

$$\sum_{k=1}^{M_n} \|\mathbf{\Gamma}^-_{k\bar{\mathcal{E}}_0}\|_1 \leq \sqrt{M_n m^2 s_{n0}}\|\mathbf{\Delta}\|_F. \quad (90)$$

Set $\|\mathbf{\Gamma}^-_k\|_1 = \|\mathbf{\Gamma}^-_{k\bar{\mathcal{E}}_0}\|_1 + \|\mathbf{\Gamma}^-_{k\bar{\mathcal{E}}_0^c}\|_1$ in A_2 of (83) to deduce that w.h.p.

$$\alpha A_2 + A_3 \geq -\alpha|A_2| + A_3$$

$$\geq \alpha(0.5\lambda_n - C_0 d_{1n}) \sum_{k=1}^{M_n} \|\mathbf{\Gamma}^-_{k\bar{\mathcal{E}}_0^c}\|_1 - \alpha(C_0 d_{1n} + \lambda_n) \sum_{k=1}^{M_n} \|\mathbf{\Gamma}^-_{k\bar{\mathcal{E}}_0}\|_1 - \alpha C_0 d_{2n} \sum_{k=1}^{M_n} \|\mathbf{\Gamma}^+_k\|_F$$

$$\geq -\alpha\left((C_0 d_{1n} + \lambda_n)\sqrt{m^2 s_{n0}} + C_0 d_{2n}\right)\sqrt{M_n}\|\mathbf{\Delta}\|_F$$

$$\geq -\alpha\left(m\sqrt{s_{n0}}\,\lambda_n + 2C_0 d_{2n}\right)\sqrt{M_n}\,\|\mathbf{\Delta}\|_F \quad (91)$$

where we have used the fact that $0.5\lambda_n \geq C_0 d_{1n} = \lambda_{n\ell}/2$ (see (46)), (90), $\sum_{k=1}^{M_n} \|\mathbf{\Gamma}^+_k\|_F \leq \sqrt{M_n}\,\|\mathbf{\Delta}\|_F$ (by the Cauchy-Schwarz inequality), and the bound $\sqrt{s_{n0}}\,md_{1n} \leq d_{2n}$. Now use A_2 of (84) to deduce that w.h.p.

$$(1-\alpha)A_2 + A_4 \geq -(1-\alpha)|A_2| + A_4$$

$$\geq (1-\alpha)m\sqrt{M_n}(0.5\lambda_n - C_0 d_{1n}) \sum_{(q,\ell)\in\mathcal{E}_0^c} \|\mathbf{\Delta}^{(q\ell M_n)}\|_F - (1-\alpha)m\sqrt{M_n}\left((C_0 d_{1n} + \lambda_n)\sum_{(q,\ell)\in\mathcal{E}_0} \|\mathbf{\Delta}^{(q\ell M_n)}\|_F + C_0 d_{1n}\sum_{q=\ell=1}^{p_n} \left\|\mathbf{\Delta}^{(q\ell M_n)}\right\|_F\right)$$

$$\geq -(1-\alpha)\sqrt{M_n}\,\|\mathbf{\Delta}\|_F \left(\sqrt{m^2 s_{n0}}\,\lambda_n + C_0 d_{1n} m\left(\sqrt{s_{n0}} + \sqrt{p_n}\right)\right)$$

$$\geq -(1-\alpha)\left(m\sqrt{s_{n0}}\,\lambda_n + C_0(1+m)d_{2n}\right)\sqrt{M_n}\,\|\mathbf{\Delta}\|_F \quad (92)$$

where we have used the facts that $0.5\lambda_n \geq C_0 d_{1n} = \lambda_{n\ell}/2$, $\sum_{(q,\ell)\in\mathcal{E}_0} \|\mathbf{\Delta}^{(q\ell M_n)}\|_F \leq \sqrt{s_{n0}}\|\mathbf{\Delta}\|_F$ and $\sum_{q=\ell=1}^{p_n} \|\mathbf{\Delta}^{(q\ell M_n)}\|_F \leq \sqrt{p_n}\|\mathbf{\Delta}\|_F$ by the Cauchy-Schwarz inequality, and the bounds $\sqrt{s_{n0}}\,md_{1n} \leq d_{2n}$ and $\sqrt{p_n}\,md_{1n} \leq md_{2n}$.

From (91) and (92), after some simplifications, we have

$$A_2 + A_3 + A_4 \geq -\left(m\sqrt{s_{n0}}\,\lambda_n + C_0(1+m)d_{2n}\right) \times \sqrt{M_n}\,\|\mathbf{\Delta}\|_F \quad (93)$$

where we used the bound $2d_{2n} \leq (1+m)d_{2n}$. By (41), (47) and (48), λ_n is chosen to satisfy

$$\lambda_n \leq \lambda_{nu1} = \frac{C_0(1+m)}{m\sqrt{s_{n0}M_n}}\, r_n. \quad (94)$$

Noting that $r_n = \sqrt{M_n}\, d_{2n}$, we have

$$A_2 + A_3 + A_4 \geq -2C_0(1+m)r_n\|\mathbf{\Delta}\|_F. \quad (95)$$

Using (73), (82) and (95), and $\|\mathbf{\Delta}\|_F = Rr_n$, we have w.h.p.

$$G(\mathbf{\Delta}) \geq \|\mathbf{\Delta}\|_F^2 \left[\frac{1}{2(\beta_{\min}^{-1} + Rr_n)^2} - \frac{2C_0(1+m)}{R}\right]. \quad (96)$$

For the given choice of N_2, $Rr_n \leq Rr_{N_2} \leq 0.1/\beta_{\min}$ for $n \geq N_2$. Also, $2C_0(1+m)/R = \beta_{\min}^2/4$ by (40). Then for $n \geq N_2$,

$$\frac{1}{2(\beta_{\min}^{-1} + Rr_n)^2} - \frac{2C_0(1+m)}{R} \geq \beta_{\min}^2\left(\frac{1}{2.42} - \frac{1}{4}\right) > 0,$$

implying $G(\mathbf{\Delta}) > 0$. This proves (49). The choice of N_3 for log-sum penalty ensures that $|[\mathbf{\Gamma}_k]_{ij}| \leq \delta_\lambda = \epsilon$ needed in (86) is satisfied w.h.p.: if $Rr_n \leq \epsilon$, then $|[\mathbf{\Gamma}_k]_{ij}| \leq \|\mathbf{\Delta}\|_F \leq Rr_n \leq \epsilon$.

The following applies to the SCAD penalty.

For SCAD Penalty: Here we address (85) differently. Using triangle inequality, we have

$$|[\tilde{\mathbf{\Phi}}_k]_{ij}| \geq |[\mathbf{\Phi}_{0k}]_{ij}| + \gamma\left(|[\mathbf{\Phi}_{0k}]_{ij}| - |[\mathbf{\Gamma}_k]_{ij}| - |[\mathbf{\Phi}_{0k}]_{ij}|\right) \geq |[\mathbf{\Phi}_{0k}]_{ij}| - |[\mathbf{\Gamma}_k]_{ij}|. \quad (97)$$

Since $|[\mathbf{\Gamma}_k]_{ij}| \leq \|\mathbf{\Delta}\|_F \leq Rr_n$, the choice $\lambda_n = \lambda_{nu2}$ implies that $\lambda_n \geq Rr_n$, satisfying $|[\mathbf{\Gamma}_k]_{ij}| \leq \lambda_n$. Therefore, $|[\tilde{\mathbf{\Phi}}_k]_{ij}| \geq |[\mathbf{\Phi}_{0k}]_{ij}| - \lambda_n$. For $n \geq N_4$, $\rho'_\lambda(|[\tilde{\mathbf{\Phi}}_k]_{ij}|) = 0$ (see (45) if $\{i,j\} \in \bar{\mathcal{E}}_0$, i.e, $[\mathbf{\Phi}_{0k}]_{ij} \neq 0$, since in this case $|[\tilde{\mathbf{\Phi}}_k]_{ij}| \geq (a +$

$1)\lambda_n - \lambda_n = a\lambda_n$. Therefore, for $n \geq N_4$,

$$\begin{aligned} A_{3k} &= \alpha \sum_{(i,j)\in\bar{\mathcal{E}}_0^c} \rho_\lambda\left([\mathbf{\Gamma}_k]_{ij}\right) \\ &\geq \alpha \sum_{(i,j)\in\bar{\mathcal{E}}_0^c} C_\lambda\,|[\mathbf{\Gamma}_k]_{ij}| \quad \text{for} \quad |[\mathbf{\Gamma}_k]_{ij}| \leq \delta_\lambda \\ &= \alpha(\lambda_n/2)\|\mathbf{\Gamma}^-_{k\bar{\mathcal{E}}_0^c}\|_1 \end{aligned} \tag{98}$$

leading to ($A_3 = \sum_{k=1}^{M_n} A_{3k}$)

$$A_3 \geq \alpha\,(\lambda_n/2)\sum_{k=1}^{M_n}\|\mathbf{\Gamma}^-_{k\bar{\mathcal{E}}_0^c}\|_1\,. \tag{99}$$

Mimicking the steps for bounding A_3 above and under the same conditions, we have

$$A_4 \geq (1-\alpha)m\sqrt{M_n}\,(\lambda_n/2)\sum_{(q,\ell)\in\mathcal{E}_0^c}\|\mathbf{\Delta}^{(q\ell M_n)}\|_F\,. \tag{100}$$

Thus w.h.p.

$$\begin{aligned} &\alpha A_2 + A_3 \geq -\alpha|A_2| + A_3 \\ &\geq \alpha(0.5\lambda_n - C_0 d_{1n})\sum_{k=1}^{M_n}\|\mathbf{\Gamma}^-_{k\bar{\mathcal{E}}_0^c}\|_1 \\ &\quad - \alpha C_0 d_{1n}\sum_{k=1}^{M_n}\|\mathbf{\Gamma}^-_{k\bar{\mathcal{E}}_0}\|_1 - \alpha C_0 d_{2n}\sum_{k=1}^{M_n}\|\mathbf{\Gamma}^+_k\|_F \\ &\geq -\alpha\left(C_0 d_{1n}\sqrt{m^2 s_{n0}} + C_0 d_{2n}\right)\sqrt{M_n}\|\mathbf{\Delta}\|_F \\ &\geq -\alpha 2 C_0 d_{2n}\sqrt{M_n}\,\|\mathbf{\Delta}\|_F \end{aligned} \tag{101}$$

where we have used (90), $\sum_{k=1}^{M_n}\|\mathbf{\Gamma}^+_k\|_F \leq \sqrt{M_n}\,\|\mathbf{\Delta}\|_F$ (by the Cauchy-Schwarz inequality), the bound $\sqrt{s_{n0}}\,md_{1n} \leq d_{2n}$, and the fact since $\lambda_n = \max(Rr_n, \lambda_{nu1})$ in Theorem 1, $0.5\lambda_n - C_0 d_{1n} \geq 0$ and therefore, the term involving $0.5\lambda_n - C_0 d_{1n}$ above can be neglected. By very similar arguments we also have

$$\begin{aligned} &(1-\alpha)A_2 + A_4 \geq -(1-\alpha)|A_2| + A_4 \\ &\geq (1-\alpha)m\sqrt{M_n}(0.5\lambda_n - C_0 d_{1n})\sum_{(q,\ell)\in\mathcal{E}_0^c}\|\mathbf{\Delta}^{(q\ell M_n)}\|_F \\ &\quad - (1-\alpha)m\sqrt{M_n}\Bigg(C_0 d_{1n}\sum_{(q,\ell)\in\mathcal{E}_0}\|\mathbf{\Delta}^{(q\ell M_n)}\|_F \\ &\quad + C_0 d_{1n}\sum_{q=\ell=1}^{p_n}\|\mathbf{\Delta}^{(q\ell M_n)}\|_F\Bigg) \\ &\geq -(1-\alpha)\sqrt{M_n}\,\|\mathbf{\Delta}\|_F\,\left(C_0 d_{1n} m(\sqrt{s_{n0}} + \sqrt{p_n})\right) \\ &\geq -(1-\alpha)C_0(1+m)d_{2n}\sqrt{M_n}\,\|\mathbf{\Delta}\|_F \end{aligned} \tag{102}$$

where we have used the facts that $0.5\lambda_n \geq C_0 d_{1n} = \lambda_{n\ell}/2$, $\sum_{(q,\ell)\in\mathcal{E}_0}\|\mathbf{\Delta}^{(q\ell M_n)}\|_F \leq \sqrt{s_{n0}}\|\mathbf{\Delta}\|_F$ and $\sum_{q=\ell=1}^{p_n}\|\mathbf{\Delta}^{(q\ell M_n)}\|_F \leq \sqrt{p_n}\|\mathbf{\Delta}\|_F$ by the Cauchy-Schwarz inequality, and the bounds $\sqrt{s_{n0}}\,md_{1n} \leq d_{2n}$ and $\sqrt{p_n}\,md_{1n} \leq md_{2n}$. From (101) and (102) we have

$$\begin{aligned} A_2 + A_3 + A_4 &\geq -C_0(1+m)d_{2n}\sqrt{M_n}\,\|\mathbf{\Delta}\|_F \\ &\geq -C_0(1+m)r_n\|\mathbf{\Delta}\|_F \end{aligned} \tag{103}$$

where we used $r_n = \sqrt{M_n}\,d_{2n}$ and the bound $2d_{2n} \leq (1+m)d_{2n}$. Mimicking (96),with probability $> 1 - 1/(mp_n)^{\tau-2}$, we have

$$\begin{aligned} G(\mathbf{\Delta}) &\geq \|\mathbf{\Delta}\|_F^2\left[\frac{1}{2(\beta_{\min}^{-1} + Rr_n)^2} - \frac{C_0(1+m)}{R}\right] \\ &\geq \beta_{\min}^2\left(\frac{1}{2.42} - \frac{1}{8}\right) > 0\,, \end{aligned} \tag{104}$$

implying $G(\mathbf{\Delta}) > 0$. This proves (49). For the SCAD penalty, we need $|[\mathbf{\Gamma}_k]_{ij}| \leq \delta_\lambda = \lambda_n$ in (98). Since $|[\mathbf{\Gamma}_k]_{ij}| \leq \|\mathbf{\Delta}\|_F \leq Rr_n$, the choice $\lambda_n = \lambda_{nu2}$ implies that $\lambda_n \geq Rr_n$, satisfying $|[\mathbf{\Gamma}_k]_{ij}| \leq \lambda_n$. This completes the proof.■

B. PROOFS OF LEMMA 1 AND THEOREMS 1 AND 3

Proof of Lemma 1: Consider $\mathcal{L}(\mathbf{\Omega}) - \frac{\mu}{2}\|\mathbf{\Omega}\|_F^2$ for some $\mu \geq 0$. By [10, Lemma 4], using Wirtinger calculus, the Hessian of $\mathcal{L}(\mathbf{\Omega})$ w.r.t.

$$\begin{aligned} \boldsymbol{y} = \big[&(\mathrm{vec}(\mathbf{\Phi}_1))^\top,\ (\mathrm{vec}(\mathbf{\Phi}_1^*))^\top,\ \cdots, \\ &(\mathrm{vec}(\mathbf{\Phi}_{M_n}))^\top,\ (\mathrm{vec}(\mathbf{\Phi}_{M_n}^*))^\top\big]^\top \in \mathbb{C}^{2m^2p_n^2M_n} \end{aligned}$$

is given by

$$\begin{aligned} \nabla^2\mathcal{L}(\mathbf{\Omega}) = \text{block-diag}\Big\{&\mathbf{\Phi}_1^{-*}\otimes\mathbf{\Phi}_1^{-1},\ \mathbf{\Phi}_1^{-1}\otimes\mathbf{\Phi}_1^{-*},\ \cdots, \\ &\mathbf{\Phi}_{M_n}^{-*}\otimes\mathbf{\Phi}_{M_n}^{-1},\ \mathbf{\Phi}_{M_n}^{-1}\otimes\mathbf{\Phi}_{M_n}^{-*}\Big\} \end{aligned} \tag{105}$$

with

$$\begin{aligned} \phi_{\min}(\nabla^2\mathcal{L}(\mathbf{\Omega})) &= \min_k \phi_{\min}^2(\mathbf{\Phi}_k^{-1}) = \frac{1}{\max_k \phi_{\max}^2(\mathbf{\Phi}_k)} \\ &= \frac{1}{\max_k \|\mathbf{\Phi}_k\|^2} \geq \beta_{\min}^2\,. \end{aligned} \tag{106}$$

Since we have $\|\mathbf{\Omega}\|_F^2 = \frac{1}{2}\boldsymbol{y}^H\boldsymbol{y}$, the Hessian of $\|\mathbf{\Omega}\|_F^2$ w.r.t. $\boldsymbol{y}$ is given by

$$\nabla^2\|\mathbf{\Omega}\|_F^2 = \boldsymbol{I}_{2m^2p_n^2M_n}\,. \tag{107}$$

It follows from (106) and (107) that $\mathcal{L}(\mathbf{\Omega}) - \frac{\mu}{2}\|\mathbf{\Omega}\|_F^2$ is positive semi-definite, hence convex, if

$$\|\mathbf{\Phi}_k\| \leq \sqrt{\frac{2}{\mu}} \quad \forall k \in [M_n]\,. \tag{108}$$

By property (v) of the penalty functions, $g(u) := \rho_\lambda(u) + \frac{\mu}{2}u^2$ is convex, for some $\mu \geq 0$, and by property (ii), it is non-decreasing on $\mathbb{R}_+$. Therefore, by the composition rules [50,

Sec. 3.2.4], $g(|[\mathbf{\Phi}_k]_{ij}|)$ and $g(\|\mathbf{\Omega}^{(q\ell M_n)}\|_F)$ are convex. Hence,

$$P_e(\mathbf{\Omega}) + \frac{\mu_e}{2}\|\mathbf{\Omega}\|_F^2 = \sum_{k=1}^{M_n}\sum_{i\neq j}^{mp}\left(\rho_\lambda(|[\mathbf{\Phi}_k]_{ij}|) + \frac{\mu_e}{2}|[\mathbf{\Phi}_k]_{ij}|^2\right) \tag{109}$$

is convex for $\mu_e = \mu \geq 0$, and similarly,

$$P_g(\mathbf{\Omega}) + \frac{\mu_g}{2}\|\mathbf{\Omega}\|_F^2 = m\sqrt{M_n}\sum_{k=1}^{M_n}\sum_{q\neq\ell}^{p}\left(\rho_\lambda(\|\mathbf{\Omega}^{(q\ell M_n)}\|_F) + \frac{\mu_g}{2m\sqrt{M_n}}\|\mathbf{\Omega}^{(q\ell M_n)}\|_F^2\right) \tag{110}$$

is convex for $\mu_g = m\sqrt{M_n}\,\mu$, where μ is the value that renders $\rho_\lambda(u) + \frac{\mu}{2}u^2$ convex. Now express $\bar{\mathcal{L}}(\mathbf{\Omega})$ as

$$\bar{\mathcal{L}}(\mathbf{\Omega}) = \alpha\bar{\mathcal{L}}_e(\mathbf{\Omega}) + (1-\alpha)\bar{\mathcal{L}}_g(\mathbf{\Omega}), \tag{111}$$

$$\bar{\mathcal{L}}_e(\mathbf{\Omega}) = \mathcal{L}(\mathbf{\Omega}) - \frac{\mu}{2}\|\mathbf{\Omega}\|_F^2 + P_e(\mathbf{\Omega}) + \frac{\mu}{2}\|\mathbf{\Omega}\|_F^2, \tag{112}$$

$$\bar{\mathcal{L}}_g(\mathbf{\Omega}) = \mathcal{L}(\mathbf{\Omega}) - \frac{\mu}{2}\|\mathbf{\Omega}\|_F^2 + P_g(\mathbf{\Omega}) + \frac{\mu}{2}\|\mathbf{\Omega}\|_F^2. \tag{113}$$

By (108), (109) and (112), $\bar{\mathcal{L}}_e(\mathbf{\Omega})$ is convex function of $\mathbf{\Omega}$ if

$$\|\mathbf{\Phi}_k\| \leq \sqrt{\frac{2}{\mu}} \quad \forall k \in [M_n], \tag{114}$$

and by (108), (110) and (113), $\bar{\mathcal{L}}_g(\mathbf{\Omega})$ is convex in $\mathbf{\Omega}$ if

$$\|\mathbf{\Phi}_k\| \leq \sqrt{\frac{2}{\mu_g}} = \sqrt{\frac{2}{m\mu\sqrt{M_n}}} \quad \forall k \in [M_n]. \tag{115}$$

Thus, for $\bar{\mathcal{L}}(\mathbf{\Omega})$ to be strictly convex, using the (minimum) values of μ to make $\rho_\lambda(u) + \frac{\mu}{2}u^2$ convex, we require

$$\|\mathbf{\Phi}_k\| < \sqrt{\frac{2}{m\mu\sqrt{M_n}}} \quad \forall k \in [M_n] = \begin{cases} \infty & : \text{Lasso} \\ \sqrt{\frac{2(a-1)}{m\sqrt{M_n}}} & : \text{SCAD} \\ \sqrt{\frac{2\epsilon}{m\sqrt{M_n}\lambda_n}} & : \text{log-sum}, \end{cases} \tag{116}$$

The choice $\|\mathbf{\Phi}_k\| < \sqrt{\frac{2}{m\mu\sqrt{M_n}}}$ makes $\mathcal{L}(\mathbf{\Omega}) - \frac{\mu}{2}\|\mathbf{\Omega}\|_F^2$ positive definite, hence strictly convex. We take $\|\mathbf{\Phi}_k\| = 0.99\sqrt{\frac{2}{m\mu\sqrt{M_n}}}$, completing the proof.■

Proof of Theorem 2: If $1/\beta_{\min} \leq 0.99\sqrt{\frac{2}{m\mu\sqrt{M_n}}}$, then $\mathbf{\Phi}_{0k} \in \mathcal{B}_k$ since $\|\mathbf{\Phi}_{0k}\| = 1/\phi_{\min}(\mathbf{S}_{0k}) \leq 1/\beta_{\min}$ by assumption (A4). To establish that $\hat{\mathbf{\Phi}}_k \in \mathcal{B}_k$, consider ($\mathbf{\Delta}$ is as in the proof of Theorem 1)

$$\begin{aligned}\|\hat{\mathbf{\Phi}}_k\| &\leq \|\hat{\mathbf{\Phi}}_k - \mathbf{\Phi}_{0k}\| + \|\mathbf{\Phi}_{0k}\| \\ &\leq \|\mathbf{\Delta}\| + 1/\beta_{\min} \leq \|\mathbf{\Delta}\|_F + 1/\beta_{\min} \\ &\leq Rr_n + 1/\beta_{\min}.\end{aligned} \tag{117}$$

Therefore, $\hat{\mathbf{\Phi}}_k \in \mathcal{B}_k$. Thus, both $\hat{\mathbf{\Phi}}_k$ and $\mathbf{\Phi}_{0k}$, hence $\hat{\mathbf{\Omega}}_\lambda$ and $\mathbf{\Omega}_0$, respectively, are feasible. The desired result then follows from Theorem 1 and (local) strict convexity of $\bar{\mathcal{L}}(\mathbf{\Omega})$ over $\cap_{k=1}^{M_n}\mathcal{B}_k$ implied by Lemma 1.■

Proof of Theorem 3: We have $\|\hat{\mathbf{\Omega}}^{(q\ell M_n)} - \mathbf{\Omega}_0^{(q\ell M_n)}\|_F \leq \|\hat{\mathbf{\Omega}} - \mathbf{\Omega}_0\|_F \leq \bar{\sigma}_n$ w.h.p. For the edge $\{q,\ell\} \in \mathcal{E}_0$, we have

$$\begin{aligned}\|\hat{\mathbf{\Omega}}^{(q\ell M_n)}\|_F &= \|\mathbf{\Omega}_0^{(q\ell M_n)} + \hat{\mathbf{\Omega}}^{(q\ell M_n)} - \mathbf{\Omega}_0^{(q\ell M_n)}\|_F \\ &\geq \|\mathbf{\Omega}_0^{(q\ell M_n)}\|_F - \|\hat{\mathbf{\Omega}}^{(q\ell M_n)} - \mathbf{\Omega}_0^{(q\ell M_n)}\|_F \\ &\geq \nu - \bar{\sigma}_n \geq 0.6\,\nu \quad \text{for} \quad n \geq N_5 \\ &> \gamma_n.\end{aligned} \tag{118}$$

Thus, $\mathcal{E}_0 \subseteq \hat{\mathcal{E}}$. Now consider the set complements $\mathcal{E}_0^c$ and $\hat{\mathcal{E}}^c$. For the edge $\{q,\ell\} \in \mathcal{E}_0^c$, $\|\mathbf{\Omega}_0^{(q\ell M_n)}\|_F = 0$. For $n \geq N_5$, w.h.p. we have

$$\begin{aligned}\|\hat{\mathbf{\Omega}}^{(q\ell M_n)}\|_F &\leq \|\mathbf{\Omega}_0^{(q\ell M_n)}\|_F + \|\hat{\mathbf{\Omega}}^{(q\ell M_n)} - \mathbf{\Omega}_0^{(q\ell M_n)}\|_F \\ &\leq 0 + \bar{\sigma}_n \leq 0.4\,\nu < \gamma_n,\end{aligned} \tag{119}$$

implying that $\{q,\ell\} \in \hat{\mathcal{E}}_0^c$. Thus, $\mathcal{E}_0^c \subseteq \hat{\mathcal{E}}^c$, hence $\hat{\mathcal{E}} \subseteq \mathcal{E}_0$, establishing $\hat{\mathcal{E}} = \mathcal{E}_0$.■

REFERENCES

[1] J. Whittaker, *Graphical Models in Applied Multivariate Statistics*. New York, NY, USA: Wiley, 1990.

[2] S. L. Lauritzen, *Graphical Models*. Oxford, U.K.: Oxford Univ. Press, 1996.

[3] P. Bühlmann and S. van de Geer, *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Berlin, Germany: Springer, 2011.

[4] N. Meinshausen and P. Bühlmann, "High-dimensional graphs and variable selection with the Lasso," *Ann. Statist.*, vol. 34, no. 3, pp. 1436–1462, 2006.

[5] O. Banerjee, L. E. Ghaoui, and A. d'Aspremont, "Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data," *J. Mach. Learn. Res.*, vol. 9, pp. 485–516, 2008.

[6] R. Dahlhaus, "Graphical interaction models for multivariate time series," *Metrika*, vol. 51, pp. 157–172, 2000.

[7] A. Jung, G. Hannak, and N. Goertz, "Graphical LASSO based model selection for time series," *IEEE Signal Process. Lett.*, vol. 22, no. 10, pp. 1781–1785, Oct. 2015.

[8] J. K. Tugnait, "Graphical modeling of high-dimensional time series," in *Proc. 52nd Asilomar Conf. Signals, Syst.*, 2018, pp. 840–844.

[9] J. K. Tugnait, "Consistency of sparse-group lasso graphical model selection for time series," in *Proc. 54th Asilomar Conf. Signals, Syst., Comput.*, 2020, pp. 589–593.

[10] J. K. Tugnait, "On sparse high-dimensional graphical model learning for dependent time series," *Signal Process.*, vol. 197 Aug. 2022, Art. no. 108539.

[11] M. J. Wainwright, *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge, U.K.: Cambridge Univ. Press, 2019.

[12] J. K. Tugnait, "Sparse-group log-sum penalized graphical model learning for time series," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 5822–5826.

[13] E. J. Candès, M. B. Wakin, and S. P. Boyd, "Enhancing sparsity by reweighted ℓ_1 minimization," *J. Fourier Anal. Appl.*, vol. 14, pp. 877–905, 2008.

[14] E. Avventi, A. G. Lindquist, and B. Wahlberg, "ARMA identification of graphical models," *IEEE Trans. Autom. Control*, vol. 58, no. 5, pp. 1167–1178, May 2013.

[15] D. Alpago, M. Zorzi, and A. Ferrante, "Identification of sparse reciprocal graphical models," *IEEE Control Syst. Lett.*, vol. 22, no. 4, pp. 659–664, Oct. 2018.

[16] J. Songsiri and L. Vandenberghe, "Topology selection in graphical models of autoregressive processes," *J. Mach. Learn. Res.*, vol. 11, pp. 2671–2705, Oct. 2010.
[17] J. You, C. Yu, J. Sun, and J. Chen, "Generalized maximum entropy based identification of graphical ARMA models," *Automatica*, vol. 141, 2022, Art. no. 110319.
[18] S. Basu and G. Michailidis, "Regularized estimation in sparse high-dimensional time series models," *Ann. Statist.*, vol. 43, no. 4, pp. 1535–1567, 2015.
[19] M. Kolar, H. Liu, and E. P. Xing, "Graph estimation from multi-attribute data," *J. Mach. Learn. Res.*, vol. 15, pp. 1713–1750, 2014.
[20] J. K. Tugnait, "Sparse-group lasso for graph learning from multi-attribute data," *IEEE Trans. Signal Process.*, vol. 69, pp. 1771–1786, 2021.
[21] G. Marjanovic and V. Solo, "Vector l_0 sparse conditional independence graphs," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 2731–2735.
[22] Z. Yue, P. Sundaram, and V. Solo, "Fast block-sparse estimation for vector networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2020, pp. 5510–5514.
[23] P. Sundaram, M. Luessi, M. Bianciardi, S. Stufflebeam, M. Hämäläinen, and V. Solo, "Individual resting-state brain networks enabled by massive multivariate conditional mutual information," *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 1957–1966, Jun. 2020.
[24] S. Zhang, B. Guo, A. Dong, J. He, Z. Xu, and S. X. Chen, "Cautionary tales on air-quality improvement in Beijing," *Proc. Roy. Soc. A*, vol. 473, 2017, Art. no. 20170457.
[25] W. Chen, F. Wang, G. Xiao, J. Wu, and S. Zhang, "Air quality of Beijing and impacts of the new ambient air quality standard," *Atmosphere*, vol. 6, pp. 1243–1258, 2015.
[26] J. Friedman, T. Hastie, and R. Tibshirani, "A note on the group lasso and a sparse group lasso," Jan. 2010, *arXiv:1001.0736v1*.
[27] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, "A sparse-group lasso," *J. Comput. Graphical Statist.*, vol. 22, pp. 231–245, 2013.
[28] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *J. Amer. Stat. Assoc.*, vol. 96, pp. 1348–1360, Dec. 2001.
[29] C. Lam and J. Fan, "Sparsistency and rates of convergence in large covariance matrix estimation," *Ann. Statist.*, vol. 37, no. 6B, pp. 4254–4278, 2009.
[30] R. J. Wolstenholme and A. T. Walden, "An efficient approach to graphical modeling of time series," *IEEE Trans. Signal Process.*, vol. 63, no. 12, pp. 3266–3276, Jun. 2015.
[31] D. Schneider-Luftman, "*p*-value combiners for graphical modelling of EEG data in the frequency domain," *J. Neurosci. Methods*, vol. 271, pp. 92–106, 2016.
[32] Y. Matsuda, "A test statistic for graphical modelling of multivariate time series," *Biometrika*, vol. 93, no. 2, pp. 399–409, 2006.
[33] J. K. Tugnait, "Edge exclusion tests for graphical model selection: Complex Gaussian vectors and time series," *IEEE Trans. Signal Process.*, vol. 67, no. 19, pp. 5062–5077, Oct. 2019.
[34] J. Songsiri, J. Dahl, and L. Vandenberghe, "Graphical models of autoregressive processes," in *Convex Optimization in Signal Processing and Communications*, Y. Eldar and D. Palomar Eds., Cambridge, U.K.: Cambridge Univ. Press, 2009, pp. 89–116.
[35] M. Eichler, "Graphical modelling of dynamic relationships in multivariate time series," in *Handbook of Time Series Analysis: Recent Theoretical Developments and Applications*, B. Schelter, M. Winterhalder, and J. Timmer, Eds., New York, NY, USA: Wiley-VCH, 2006, pp. 335–372.
[36] M. Eichler, "Graphical modelling of multivariate time series," *Probability Theory Related Fields*, vol. 153, no. 1/2, pp. 233–268, Jun. 2012.
[37] J. K. Tugnait, "Graph learning from multivariate dependent time series via a multi-attribute formulation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2022, pp. 4508–4512.
[38] J. K. Tugnait, "Sparse-group non-convex penalized multi-attribute graphical model selection," in *Proc. 29th Eur. Signal Process. Conf.*, 2021, pp. 1850–1854.
[39] H. Zou and R. Li, "One-step sparse estimates in nonconcave penalized likelihood models," *Ann. Statist.*, vol. 36, no. 4, pp. 1509–1533, 2008.
[40] J. K. Tugnait, "Conditional independence graph estimation from multi-attribute dependent time series," in *Proc. IEEE 34th Int. Workshop Mach. Learn. Signal Process.*, 2024, pp. 1–6.
[41] P.-L. Loh and M. J. Wainwright, "Support recovery without incoherence: A case for nonconvex regularization," *Ann. Statist.*, vol. 45, pp. 2455–2482, 2017.
[42] D. R. Brillinger, *Time Series: Data Analysis and Theory*, Expanded ed. New York, NY, USA: McGraw Hill, 1981.
[43] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. Roy. Stat. Soc. Ser. B, Stat. Methodol.*, vol. 68, no. 1, pp. 49–67, 2006.
[44] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, 2010.
[45] B. Zhao, Y. S. Wang, and M. Kolar, "FuDGE: A method to estimate a functional differential graph in a high-dimensional setting," *J. Mach. Learn. Res.*, vol. 23, pp. 1–82, 2022.
[46] P. Ravikumar, M. J. Wainwright, G. Raskutti, and B. Yu, "High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence," *Electron. J. Statist.*, vol. 5, pp. 935–980, 2011.
[47] R. S. Tsay, *Analysis of Financial Time Series*, 3rd ed. Hoboken, NJ, USA: Wiley, 2010.
[48] A. J. Rothman, P. J. Bickel, E. Levina, and J. Zhu, "Sparse permutation invariant covariance estimation," *Electron. J. Statist.*, vol. 2, pp. 494–515, 2008.
[49] P. J. Schreier and L. L. Scharf, *Statistical Signal Processing of Complex-Valued Data*. Cambridge, U.K.: Cambridge Univ. Press, 2010.
[50] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[51] H. Liu, K. Roeder, and L. Wasserman, "Stability approach to regularization selection (StARS) for high dimensional graphical models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 1432–1440.
[52] W. Wang, M. J. Wainwright, and K. Ramchandran, "Information-theoretic bounds on model selection for Gaussian Markov random fields," in *Proc. IEEE Int. Symp. Inf. Theory*, Austin, TX, USA, 2010, pp. 1373–1377.
[53] G. Hannak, A. Jung, and N. Görtz, "On the information-theoretic limits of graphical model selection for Gaussian time series," in *Proc. 22nd Eur. Signal Process. Conf.*, Lisbon, Portugal, 2014, pp. 516–520.

JITENDRA K. TUGNAIT (Life Fellow, IEEE) received the B.Sc. (Hons.) degree in electronics and electrical communication engineering from Punjab Engineering College, Chandigarh, India, in 1971, the M.S. and the E.E. degrees in electrical engineering from Syracuse University, Syracuse, NY, USA, and the Ph.D. degree in electrical engineering from the University of Illinois Urbana-Champaign, Champaign, IL, USA, in 1973, 1974, and 1978, respectively. From 1978 to 1982, he was an Assistant Professor of electrical and computer engineering with the University of Iowa, Iowa City, IA, USA. From 1982 to 1989, he was with the Long Range Research Division of the Exxon Production Research Company, Houston, TX, USA. In 1989, he joined the Department of Electrical and Computer Engineering, Auburn University, Auburn, AL, USA, as a Professor, where he is currently James B. Davis Professor. His research interests include statistical signal processing and machine learning for signal processing. Dr. Tugnait was an Associate Editor for the IEEE Transactions on Automatic Control, IEEE Transactions on Signal Processing, IEEE Signal Processing Letters, and IEEE Transactions on Wireless Communications, and the Senior Area Editor of the IEEE Transactions on Signal Processing, and the Senior Editor of IEEE Wireless Communications Letters.